DISCRETE AND CONTINUOUS
DYNAMICAL SYSTEMS SERIES S
Volume **15**, Number **4**, April **2022**

doi:10.3934/dcdss.2021103

pp. **747–771**

ISALT: INFERENCE-BASED SCHEMES ADAPTIVE TO LARGE TIME-STEPPING FOR LOCALLY LIPSCHITZ ERGODIC SYSTEMS

Xingjie Helen Li

Department of Mathematics and Statistics, University of North Carolina at Charlotte
9201 Univ City Blvd., Charlotte, NC 28023, USA

Fei Lu*

Department of Mathematics, Johns Hopkins University
3400 N. Charles Street, Baltimore, MD 21218, USA

Felix X.-F. Ye

Department of Mathematics and Statistics, State University of New York at Albany
Earth Science 110, 1400 Washington Avenue, Albany, NY 12222, USA

Abstract. Efficient simulation of SDEs is essential in many applications, particularly for ergodic systems that demand efficient simulation of both short-time dynamics and large-time statistics. However, locally Lipschitz SDEs often require special treatments such as implicit schemes with small time-steps to accurately simulate the ergodic measures. We introduce a framework to construct inference-based schemes adaptive to large time-steps (ISALT) from data, achieving a reduction in time by several orders of magnitudes. The key is the statistical learning of an approximation to the infinite-dimensional discrete-time flow map. We explore the use of numerical schemes (such as the Euler-Maruyama, the hybrid RK4, and an implicit scheme) to derive informed basis functions, leading to a parameter inference problem. We introduce a scalable algorithm to estimate the parameters by least squares, and we prove the convergence of the estimators as data size increases.

We test the ISALT on three non-globally Lipschitz SDEs: the 1D double-well potential, a 2D multiscale gradient system, and the 3D stochastic Lorenz equation with a degenerate noise. Numerical results show that ISALT can tolerate time-step magnitudes larger than plain numerical schemes. It reaches optimal accuracy in reproducing the invariant measure when the time-step is medium-large.

1. **Introduction.** Efficient and accurate simulation of SDEs is important in many applications such as Monte Carlo sampling, data assimilation and predictive modeling (see e.g., [3, 4, 16, 17, 19, 31]). In particular, ergodic stochastic systems often demand efficient and accurate simulation of both short-time dynamics and large-time statistics [34]. Explicit schemes, while efficient and accurate for short-time,

2020 *Mathematics Subject Classification.* Primary: 65C30, 60H35; Secondary: 37M25, 62M20.
Key words and phrases. Stochastic differential equations, inference-based scheme, model reduction in time, locally Lipschitz ergodic systems, data-driven modeling.
XL is supported by NSF DMS CAREER-1847770. FL is supported NSF DMS 1913243 and NSF DMS 1821211. FY is supported by AMS-Simons travel grants.
* Corresponding author: Fei Lu.

tend to miss the invariant measure in large-time simulations because of the accumulation of numerical error. In particular, for locally Lipschitz SDEs, they tend to be numerical unstable and may miss the invariant measure even for the small time-step (for example, the Euler-Maruyama scheme, because it destroys the Lyapunov structure [34, 36]) and require special treatments such as taming schemes under small time-step size [11, 12]. Implicit schemes, on the other hand, are numerically stable and can accurately simulate the invariant measure when the time-step is small. However, they are computationally inefficient due to the limited time-step size and the costly implicit step.

We introduce ISALT, inference-based schemes adaptive to large time-stepping, a statistical learning framework to construct explicit schemes with large time-steps that can exceed the accuracy or stability threshold of classical numerical schemes, particularly for non-globally Lipschitz systems. ISALT infers parametric explicit schemes from data generated by implicit schemes, thus it inherits the implicit schemes' accuracy in producing the invariant measure while maintaining the efficiency of explicit schemes. The inference is done once for all and the inferred scheme can be used for general purpose simulations, either a long trajectory or ensembles of short trajectories with different initial distributions, in applications such as data assimilation and uncertainty quantification [3, 18].

More specifically, we seek large time-step approximations of the ergodic SDE with additive noise

$$d\mathbf{X}_t = f(\mathbf{X}_t)dt + \sigma d\mathbf{B}_t, \tag{1.1}$$

where the drift $f : \mathbb{R}^d \to \mathbb{R}^d$ is local-Lipschitz. Here $\mathbf{B}$ is a standard m-dimensional Brownian motion with $m \leqslant d$, the diffusion matrix $\sigma \in \mathbb{R}^{d\times m}$ has linearly independent columns, and they represent a degenerate noise when $m < d$. Our goal is to design an explicit scheme with large time-stepping so that it can efficiently and accurately simulate both short-time dynamics and long-time statistics such as the invariant measure.

We infer such explicit schemes with large time-stepping from offline data generated by an implicit scheme. Figure 1 shows the schematic plot of the procedure. The essential task is to approximate the infinite-dimensional discrete-time flow map. A major difficulty in a statistical learning approach is the curse of dimensionality (COD) when using generic basis functions. Our key contribution is to approximate the flow map by parametrization of numerical schemes, which provides informed basis functions, thus avoiding the COD by harnessing the rich information and structure in classical numerical schemes. We also introduce a scalable algorithm to compute the maximal likelihood estimator by least squares, which is asymptotically normal as the data size increases (see Theorem 3.5). Furthermore, we show that the inferred scheme, when it is a parametrization of an explicit scheme and when the data size is large, has the same 1-step strong order as the explicit scheme.

In this study, we focus on learning approximate flow maps that use only the increments of the Brownian motion on each time interval (that is, the function $F^\delta(X_{t_n}, \Delta B_{t_n})$ in Figure 1). We explore the derivation of informed-basis functions from three types of classical numerical schemes: the Euler-Maruyama (EM) [17], the hybrid RK4 (fourth-order Runge-Kutta) [7], and the implicit stochastic split backward Euler (SSBE) [34], and we denote the inferred schemes by IS-EM, IS-RK4 and IS-SSBE. We test them on three non-globally Lipschitz SDEs: the 1D double-well potential, a 2D multiscale gradient system, and the 3D stochastic Lorenz equation

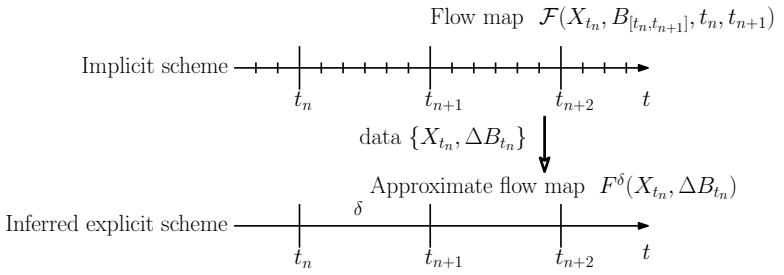


FIGURE 1. Schematic plot of inferring explicit scheme with a large time-step.

with degenerate noise. Numerical results show that the inferred schemes can tolerate time-steps ten to hundreds times larger than the plain numerical schemes, and they reach optimal accuracy in reproducing the invariant measures at medium large time-steps (see Figures 2, 4, 7, and 8). Overall, IS-RK4 produces the most accurate invariant measures in all examples, particularly when the dynamics is dominated by the drift (e.g., the Lorentz system) because the RK4 provides a higher order approximation to the drift.

Discretization with large time-stepping for differential equations (SDEs, ODEs and PDEs) is a model reduction in time, part of the general problems of space-time model reduction (see e.g., [2, 4, 10, 16, 19, 21, 26, 32]). Since the large time-step prevents classical numerical approximations based on Taylor expansions, data-driven approaches have been the primary efforts and have witnessed many successes, including the time series approaches (see e.g., [2, 24, 28]) and deep learning methods that can efficiently solve high-dimensional PDEs and SDEs on rough space-time meshes (see e.g., [1, 6, 25, 38, 39]), to name just a few. In these approaches, the discrete-time models account for the effects of the unresolved dynamics in an averaged fashion through inference, which lead to computationally efficient models for the effective dynamics [2, 19, 20, 27]. The contribution of our ISALT is to provide a simple yet effective approach to achieve large time-stepping by combining inference with classical numerical schemes. In particular, the explicit parametric form in ISALT clearly identifies the connection between classical numerical scheme and the model inferred from data. It provides a ground for further understanding the fundamental issues of data-based reduced models, such as quantification of the approximation and optimality of the reduction in time or in space-time.

The exposition of our study proceeds as follows. We first summarize the notations in Table 1. After introducing a flow map view for numerical schemes, we introduce in Section 2 the ISALT framework, that is, the procedure and algorithm for inferring schemes adaptive to the large time-step from data. Section 3 presents the theoretical results on the convergence of the estimators. In Section 4, we test ISALT on the three typical non-globally Lipschitz SDEs. Section 5 concludes our main findings with an outlook of future research.

2. **Inference of explicit schemes from data.**

Throughout this study, we assume that the SDE (1.1) is ergodic. Roughly speaking, a sufficient condition (see [15, 34]) for the SDE to be ergodic) is when (i) there is a Lyapunov function ensuring global stability and (ii) the SDE satisfies a minority condition that ensures recurrence.

Our goal is to design a numerical scheme with a large time-step, which can exceed the accuracy or stability threshold of classical numerical schemes, so that

TABLE 1. Notations

Notation	Description
$\mathbf{X}_t$ and $\mathbf{B}_t$	true state process and original stochastic force
$f(\mathbf{X}_t)$, $\sigma \in \mathbb{R}^{d\times m}$	local-Lipschitz drift and diffusion matrix
dt	time-step generating data
$\delta = \text{Gap} \times dt$	time-step for inferred scheme, $\text{Gap} \in \{1, 2, 4, 10, 20, 40, \ldots\}$
$t_i = i\delta$	discrete time instants of data
$\{\mathbf{X}^{(m)}_{t_0:t_N}, \mathbf{B}^{(m)}_{t_0:t_N}\}_{m=1}^M$	Data: M independent paths of $\mathbf{X}$ and $\mathbf{B}$ at discrete-times
$\mathcal{F}\left(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1})}\right)$	true flow map representing $(\mathbf{X}_{t_{i+1}} - \mathbf{X}_{t_i})/\delta$
$F^\delta(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})$	approximate flow map using only $\mathbf{X}_{t_n}$, $\Delta\mathbf{B}_{t_n} = \mathbf{B}_{t_{n+1}} - \mathbf{B}_{t_n}$
$\widetilde{F}^\delta\left(c^\delta, \mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n}\right)$	parametric approximate flow map
$c^\delta = (c^\delta_0, \ldots, c^\delta_p)$	parameters to be estimated for the inferred scheme
η_n and σ^δ_η	iid $N(0, I_d)$ and covariance, representing regression residual
EM and IS-EM	Euler-Maruyama and inferred scheme (IS) parametrizing it
HRK4 and IS-RK4	hybrid RK4 and inferred scheme parametrizing RK4
SSBE and IS-SSBE	split-step stochastic backward Euler and IS parametrizing it

it can efficiently and accurately simulate both short-time dynamics and long-time statistics such as invariant measures. This is of particular interest for SDEs with non-globally Lipschitz drift, because explicit schemes such as Euler-Maruyama often blow up or miss the invariant measures even if they are stable [34, 36] and implicit schemes are computationally costly while being accurate in long-time statistics.

We obtain explicit schemes with large time-steps through inference from offline data generated by an implicit scheme. The key is to approximate the flow map by parametrization of numerical schemes, instead of using a generic basis, to avoid the curse of dimensionality in the statistical learning of the flow map. Toward the goal, we will first introduce the view that numerical schemes are approximations of the flow map, then we outline the framework of statistical learning of the flow map.

2.1. **A flow map view of numerical schemes.** A numerical scheme aims to approximate the discrete-time flow map of the stochastic process. More precisely, for a time-step $\delta > 0$, let $t_i = i\delta$ and denote $(\mathbf{X}_{t_i}, i \geqslant 0)$ the process defined in (1.1) at discrete times. Based on the Markov property of $(\mathbf{X}_t)$, a numerical scheme approximates the flow map

$$\begin{aligned}\mathbf{X}_{t_{i+1}} - \mathbf{X}_{t_i} = \int_{t_i}^{t_{i+1}} f(\mathbf{X}_s)ds + \int_{t_i}^{t_{i+1}} \sigma d\mathbf{B}_s &= \delta\mathcal{F}(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1}) \\ &\approx \delta F^\delta(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}),\end{aligned} \tag{2.1}$$

where $\mathcal{F}$ is a functional depending on $\mathbf{X}_{t_i}$, the continuous trajectory $\mathbf{B}_{[t_i,t_{i+1}]}$, t_i, and t_{i+1}. The simplest scheme approximates the functional by a function $F^\delta(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ on $\mathbb{R}^{2d}$, in which one represents $\mathbf{B}_{[t_i,t_{i+1}]}$ by its increment on the time interval $\Delta\mathbf{B}_{t_i} = \mathbf{B}_{t_{i+1}} - \mathbf{B}_{t_i} \sim \mathcal{N}(0, \delta I_d)$. Among many such schemes, for example ([9, 17, 22, 33, 34]), we consider three simple and representative examples: the explicit Euler-Maruyama scheme (EM) scheme [17], the hybrid RK4 (HRK4) [7],

and the split-step stochastic backward Euler (SSBE) [34]

$$\begin{aligned} &\text{EM} \quad \mathbf{X}_{n+1} = \mathbf{X}_n + f(\mathbf{X}_n)\delta + \sigma\Delta\mathbf{B}_n, \\ &\text{HRK4} \quad \mathbf{X}_{n+1} = \mathbf{X}_n + \phi_1^{RK4}(\mathbf{X}_n, \sigma\Delta\mathbf{B}_n)\delta + \sigma\Delta\mathbf{B}_n, \\ &\text{SSBE} \quad \mathbf{X}_{n+1} = \mathbf{X}_* + \sigma\Delta\mathbf{B}_n, \text{ with } \mathbf{X}_* = \mathbf{X}_n + f(\mathbf{X}_*)\delta, \end{aligned} \tag{2.2}$$

where the term ϕ_1^{RK4} is a standard RK4 step with the stochastic force treated as a constant input:

$$\begin{aligned} \phi_1^{RK4}(\mathbf{X}_n, \sigma\Delta\mathbf{B}_n) := (k_1 + 2k_2 + 2k_3 + k_4)/6, \text{ with} \\ k_1 = f(\mathbf{X}_n) + \sigma\Delta\mathbf{B}_n/\delta, \\ k_2 = f(\mathbf{X}_n + k_1 \cdot \delta/2) + \sigma\Delta\mathbf{B}_n/\delta, \\ k_3 = f(\mathbf{X}_n + k_2 \cdot \delta/2) + \sigma\Delta\mathbf{B}_n/\delta, \\ k_4 = f(\mathbf{X}_n + k_3 \cdot \delta/2) + \sigma\Delta\mathbf{B}_n/\delta. \end{aligned}$$

Correspondingly, they approximate the flow map $\mathcal{F}(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1})$ by

$$\begin{aligned} &\text{EM} \quad F_{EM}^{\delta}(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = f(\mathbf{X}_{t_i}) + \sigma\Delta\mathbf{B}_{t_i}/\delta, \\ &\text{HRK4} \quad F_{RK4}^{\delta}(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = \phi_1^{RK4}(\mathbf{X}_{t_i}, \sigma\Delta\mathbf{B}_{t_i}) + \sigma\Delta\mathbf{B}_{t_i}/\delta, \\ &\text{SSBE} \quad F_{SSBE}^{\delta}(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = (\mathbf{X}_* - \mathbf{X}_{t_i})/\delta + \sigma\Delta\mathbf{B}_{t_i}/\delta, \end{aligned} \tag{2.3}$$

where $\mathbf{X}_* = \mathbf{X}_{t_i} + f(\mathbf{X}_*)\delta$.

For short time simulation, these schemes are of strong order 1, i.e., the discrete approximations converge to the true solution trajectory-wisely in probability at order $O(\delta)$ as the time-step vanishes, since the noise is additive [9, 17, 34, 37]. For large time simulation aiming to approximate the invariant measure, the explicit schemes can be problematic for local Lipschitz drifts and degenerate noises, for instance, the EM scheme may destroy the Lyapunov structure and fail to be ergodic for any choice of time-step [34, Lemma 6.3]. The implicit scheme SSBE, on the other hand, is ergodic and produces accurate invariant measure when the time-step is sufficiently small [34, Section 6].

In many applications, it is desirable to have an efficient numerical scheme being accurate in both short-time and large-time. A drawback of an implicit scheme is its inefficiency: it has to solve a fixed point problem in the implicit step, which is computationally costly and limits the time-step size. Taking advantage of implicit schemes, we use them to generate data and learn efficient explicit schemes with large time-steps from the data.

2.2. **Inference of a scheme from data.** We infer from data an explicit scheme that is accurate in both short-time dynamics and large-time statistics. It maintains the efficiency of explicit schemes while preserving the invariant measure as implicit schemes. The key idea is to learn an approximation of the flow map from data. To avoid the curse of dimensionality in the learning of the flow map, which is often high-dimensional and nonlinear, we derive parametric functions from the system and its numerical schemes. Roughly, the inference consists of four parts:

1. Generation of faithful data by an implicit scheme with a small time-step;
2. Derivation of a parametric form to approximate the flow map, by extracting basis functions from the system and its numerical approximations;
3. Parameter estimation by maximal likelihood methods, which leads to a least-squares problem when the parametric form is linear in the parameters;

4. Model selection: by cross-validation and convergence criteria.

Data generation. We generate faithful data, consisting of trajectories of the process at discrete times $\{t_i = i\delta\}$, by an accurate implicit scheme. That is, we first solve the system by an implicit scheme with a small time-step $\Delta t < \delta$, then we downsample the solution at the discrete times. We also save the trajectory data of the stochastic force $(\mathbf{B}_t)$. Denote these trajectories by

$$\text{Data: } \{\mathbf{X}^{(m)}_{t_0:t_N}, \mathbf{B}^{(m)}_{t_0:t_N}\}_{m=1}^{M}, \tag{2.4}$$

where N denotes the number of observing time grids and M denotes the number of independent trajectories.

The initial conditions $\{\mathbf{X}^{(m)}_{t_0}\}_{m=1}^{M}$ are samples from either a long trajectory, which represents the invariant measure, or an initial distribution that helps to explore the distribution of the process.

Derivation of parametric form. The major difficulty in inference is the approximation of the flow map $\mathcal{F}(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1})$, which is an infinite-dimensional functional. When using a non-parametric approach with the generic dictionary or basis functions, one encounters the well-known curse-of-dimensionality (COD): the size of the dictionary or basis functions increases exponentially as the dimension increases. Recent efforts on overcoming the COD include selecting adaptive-to-data basis functions in a nonparametric fashion [14], assuming a low-dimensional interaction between the components of the state variable in the spirit of particle interactions [30], or deep learning methods that approximate high dimensional functions through compositions of simple functions [1, 6, 25, 38, 39].

We take a semi-parametric approach: we avoid the COD by deriving parametric functions from the full system and its numerical schemes, which provide rich information about the flow map. In particular, we aim for parametric functions depending linearly on the parameters, so that the parameters can be estimated by least squares and our algorithm is scalable.

We focus on approximating the flow map $\mathcal{F}(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1})$ by the simplest functions $F^\delta(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$, in a parametric form

$$F^\delta(c^\delta; x, \xi) = \sum_{i=0}^{p} c_i^\delta \phi_i(x, \xi), \tag{2.5}$$

with ξ having the same distribution as $\Delta\mathbf{B}_{t_i}$. Here $\phi_i : \mathbb{R}^{2d} \to \mathbb{R}^d$ are basis functions to be extracted from numerical schemes (see Section 2.3), and $\{c_i^\delta\}$ are the parameters to be estimated from data. That is, with $(\mathbf{X}_n, \xi_n)$ corresponding to $(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})$, we infer the following scheme

$$\mathbf{X}_{n+1} = \mathbf{X}_n + \delta F^\delta(c^\delta; \mathbf{X}_n, \xi_n) + \delta\sigma_\eta\eta_n = \mathbf{X}_n + \delta\sum_{i=0}^{p} c_i^\delta\phi_i(\mathbf{X}_n, \xi_n) + \delta\sigma_\eta^\delta\eta_n, \tag{2.6}$$

where we add $\{\sigma_\eta^\delta\eta_n\}$ to account for the residual of the regression. This additional noise term can be important for the reproduction of the invariant measure. For convenience, we assume that $\{\eta_n\}$ is a sequence of iid Gaussian $N(0, I_d)$ random variables and is independent of $\{\xi_n\}$, and σ_η^δ is a diagonal matrix. In general, this residual term can be colored noise or multiplicative noise.

In view of statistical learning, the function (2.5) approximates the flow map in the function space $\mathcal{H} = \text{span}\{\phi_i(x, \xi)\}_{i=0}^{p}$, which is a subspace of $L^2(\mathbb{R}^{2d}, \mu \otimes \nu)$ with μ being the invariant measure of $\mathbf{X}$ and $\nu \sim \mathcal{N}(0, \delta I_d)$ being the distribution

of ξ (which represents $\Delta\mathbf{B}_{t_i}$). We refer $\{\phi_i(x,\xi)\}$ as basis functions and will extract them from numerical schemes (see Section 2.3).

Here we focus on using only $\Delta\mathbf{B}_{t_n}$, but one can use more sample points of the trajectory $\mathbf{B}_{[t_n,t_{n+1}]}$ and extract terms from high-order approximations based on multiple stochastic integral [9]. We postpone this as future work.

Parameter estimation. We estimate the parameters by maximizing the likelihood for the model in (2.6) with the data $\{\mathbf{X}^{(m)}_{t_0:t_N},\mathbf{B}^{(m)}_{t_0:t_N}\}_{m=1}^M$:

$$l(c^\delta_{0:p}) = \frac{1}{M}\sum_{m=1}^{M} l(\mathbf{X}^{(m)}_{t_0:t_N},\mathbf{B}^{(m)}_{t_0:t_N} \mid c^\delta_{0:p})$$

with $l(\mathbf{X}^{(m)}_{t_0:t_N},\mathbf{B}^{(m)}_{t_0:t_N} \mid c^\delta_{0:p})$ denoting the likelihood of the m-th trajectory:

$$\begin{aligned} &l(\mathbf{X}_{t_0:t_N},\mathbf{B}_{t_0:t_N} \mid c^\delta_{0:p})\\ =&\frac{1}{N}\sum_{k=1}^{d}\sum_{n=0}^{N-1}\left[\frac{|\mathbf{X}^k_{t_{n+1}}-\mathbf{X}^k_{t_n}-\delta F^\delta_k(c^\delta,\mathbf{X}_{t_i},\Delta\mathbf{B}_{t_n})|^2}{2\sigma^2_{k,\delta}}-\frac{1}{2}\log(2\pi(\sigma_{k,\delta})^2)\right], \end{aligned}$$

where F^δ_k is the k-th entry of the $\mathbb{R}^d$-valued function F^δ defined in (2.5):

$$F^\delta_k(c^\delta,\mathbf{X}_{t_i},\Delta\mathbf{B}_{t_n}) = \sum_{i=0}^{p} c^\delta_{i,k}\phi^k_i(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n}).$$

Noticing that the likelihood function is quadratic in the parameters $\{c^\delta_{i,k}\}_{i=0}^p$, we estimate them by least squares:

$$\begin{aligned} \widehat{c^{\delta,N,M}_{0:p,k}} &= (\bar{A}^{N,M}_k)^+\bar{b}^{N,M}_k,\\ (\widehat{\sigma^{\delta,N,M}_{\eta,k}})^2 &= \frac{1}{N}\sum_{n=0}^{N-1}\left|\frac{\mathbf{X}^k_{t_{n+1}}-\mathbf{X}^k_{t_n}}{\delta}-F^\delta_k(\widehat{c^{\delta,N,M}},\mathbf{X}_{t_i},\Delta\mathbf{B}_{t_n})\right|^2, \end{aligned} \tag{2.7}$$

where A^+ denotes the pseudo-inverse of A, and the normal matrix $\bar{A}^{N,M}_k$ and vector $\bar{b}^{N,M}_k$ are given by

$$\begin{aligned} \bar{A}^{N,M}_k(i,j) &= \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}\phi^k_i(\mathbf{X}^{k,(m)}_{t_n},\Delta\mathbf{B}^{k,(m)}_{t_n})\phi^k_j(\mathbf{X}^{k,(m)}_{t_n},\Delta\mathbf{B}^{k,(m)}_{t_n}),\\ \bar{b}^{N,M}_k(i) &= \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}\frac{\mathbf{X}^{k,(m)}_{t_{n+1}}-\mathbf{X}^{k,(m)}_{t_n}}{\delta}\phi^k_i(\mathbf{X}^{k,(m)}_{t_n},\Delta\mathbf{B}^{k,(m)}_{t_n}) \end{aligned} \tag{2.8}$$

for $i,j=0,\dots,p$. Here $\widehat{\sigma^{\delta,N,M}_{\eta,k}}$, the square root of the regression's residuals, provides the diagonal entries of σ^δ_η.

The above least square regression is based on the assumption that the residual $\sigma^\delta\eta_n$ defined in (2.6) is Gaussian with uncorrelated entries. The entry-wise regression aims to reflect the dynamical scale difference between entries. One may improve the approximation by considering correlated entries or other distributions for the residual.

Model selection. The parametric form in Eq.(2.6) has many degrees of freedom underdetermined, particularly when we have multiple options for the parametric form, along with possible overfitting and redundancy in these options. We select the estimated scheme by the following criteria:

- Cross validation: the estimated scheme should be stable and can reproduce the distribution of the process, particularly the main statistics. We consider the marginal invariant densities and temporal correlations (of $(\mathbf{X}_t^k)$) :

$$\begin{aligned} p_k(z)dz &= \mathbb{E}[\mathbf{1}_{(z,z+dz)}(\mathbf{X}_{t_n}^k)] \approx \frac{1}{NM}\sum_{m,n=1}^{M,N} \mathbf{1}_{(z,z+dz)}(\mathbf{X}_{t_n}^{k,(m)}), \\ C_k(h) &= \mathbb{E}[\mathbf{X}_{t_n+h}^k \mathbf{X}_{t_n}^k] \approx \frac{1}{NM}\sum_{m,n=1}^{M,N} \mathbf{X}_{t_n+h}^{k,(m)} \mathbf{X}_{t_n}^{k,(m)} \end{aligned} \tag{2.9}$$

  for $k = 1, \dots, d$.
- Convergence of the estimators. If the model is perfect and the data are either independent trajectories or a long trajectory of a stationary process, the estimators converge to the true values when the data size increases (see Theorem 3.2). While our parametric model is not perfect, the estimators also converge when the data size increases (see Theorem 3.5).

2.3. **Parametrization of numerical schemes.** We derive parametric forms to approximate the flow map from numerical schemes. The numerical schemes provide informed basis functions for inference because of their error-controlled approximations to the flow map $\mathcal{F}(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1})$ in (2.1). These basis functions can either be simply the terms in an explicit scheme or terms approximating the implicit schemes. One may view this approach as a parametrization of numerical schemes.

We focus on using only $\Delta\mathbf{B}_{t_i}$, the increment of $\mathbf{B}_{[t_i,t_{i+1}]}$, and seek parametric functions $F^\delta(c^\delta, \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ (as in (2.5)) to approximate the flow map. This constraint has two advantages: first, it makes the inferred-scheme computationally efficient, because the inferred scheme will generate only two random numbers (ξ_i, η_i in (2.6)) in each time step to represent the stochastic forces and residuals; second, it significantly reduces the function space of inference, from a functional depending on the path $\mathbf{B}_{[t_i,t_{i+1}]}$ to a function depending only on the increments. By starting from this simple setting, we hope to provide insight on the future design of schemes using multi-point noise by parametrizing high-order stochastic schemes (see e.g. [9, 13, 17]).

The flow maps (2.3) of the numerical schemes in (2.2) provide three representative candidates for a parametric function $\widetilde{F}^\delta(c^\delta, \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$. The EM is an explicit one-step scheme, the RK4 is an explicit multi-step scheme, and the SSBE is an implicit one-step scheme. Linearly parametrizing them or their Itô-Taylor expansions, i.e., adding coefficients to the terms, we obtain parametric flow maps:

$$\begin{aligned} &\text{EM} && \widetilde{F}_{EM}^\delta(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta f(\mathbf{X}_{t_i}) + c_2^\delta \sigma \Delta\mathbf{B}_{t_i}/\delta, \\ &\text{HRK4} && \widetilde{F}_{RK4}^\delta(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta \phi_1^{RK4}(\mathbf{X}_{t_i}, \sigma\Delta\mathbf{B}_{t_i}) + c_2^\delta \sigma \Delta\mathbf{B}_{t_i}/\delta, \\ &\text{SSBE} && \widetilde{F}_{SSBE}^\delta(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta \phi_1^{SSBE}(\mathbf{X}_{t_i}) + c_2^\delta \sigma \Delta\mathbf{B}_{t_i}/\delta, \end{aligned} \tag{2.10}$$

where the function ϕ_1^{RK4} is introduced in (2.2) and ϕ_1^{SSBE} is given by

$$\phi_1^{SSBE}(\mathbf{X}_{t_i}) = (I_d - \delta \nabla f(\mathbf{X}_{t_i}))^{-1} f(\mathbf{X}_{t_i}). \tag{2.11}$$

These terms are derived as follows.

- The parametric flow maps $\widetilde{F}_{EM}^\delta(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ and $\widetilde{F}_{RK4}^\delta(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ come simply by adding coefficients to each term in F_{EM}^δ and F_{RK4}^δ of the Euler and RK4 schemes in (2.3).

- We introduced an extra linear term $c_0^\delta \mathbf{X}_{t_i}$. When f is nonlinear, it serves as a linear basis function, and it helps to data-adaptively adjust the linear stability of the inferred scheme.
- The parametric flow maps $\widetilde{F}^\delta_{SSBE}(c^\delta; \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ comes from parametrizing the terms in an approximation of $F^\delta_{SSBE}(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i})$ in (2.3). More precisely, by the mean-value theorem, there exists a state $\widetilde{\mathbf{X}}_{t_i}$ depending on $\mathbf{X}_*$ and $\mathbf{X}_{t_i}$ such that

$$\begin{aligned} f(\mathbf{X}_*) &= f(\mathbf{X}_{t_i}) + \nabla f(\widetilde{\mathbf{X}}_{t_i})(\mathbf{X}_* - \mathbf{X}_{t_i}) \\ &= f(\mathbf{X}_{t_i}) + \nabla f(\mathbf{X}_{t_i})(\mathbf{X}_* - \mathbf{X}_{t_i}) + R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f), \end{aligned} \tag{2.12}$$

  where $R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f) = [\nabla f(\widetilde{\mathbf{X}}_{t_i}) - \nabla f(\mathbf{X}_{t_i})](\mathbf{X}_* - \mathbf{X}_{t_i})$. Then, by the definition of $\mathbf{X}_*$ in the SSBE in (2.3), we have

$$\mathbf{X}_* = \mathbf{X}_{t_i} + \delta f(\mathbf{X}_*) = \mathbf{X}_{t_i} + \delta[f(\mathbf{X}_{t_i}) + \nabla f(\mathbf{X}_{t_i})(\mathbf{X}_* - \mathbf{X}_{t_i})] + R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f)$$

$$\Rightarrow (\mathbf{X}_* - \mathbf{X}_{t_i}) = (I_d - \delta\nabla f(\mathbf{X}_{t_i}))^{-1}\delta f(\mathbf{X}_{t_i}) + R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f).$$

  Thus, we have

$$F^\delta_{SSBE}(\mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_i}) = (I_d - \delta\nabla f(\mathbf{X}_{t_i}))^{-1} f(\mathbf{X}_{t_i}) + \sigma\Delta\mathbf{B}_{t_i}/\delta + R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f).$$

  Assuming that $R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f)$ is negligible, parametrizing the other terms, and adding $c_0^\delta \mathbf{X}_{t_i}$, we obtain $\widetilde{F}^\delta_{SSBE}$ with ϕ_1^{SSBE} above. Note that when f is globally Lipschitz (thus $|\nabla f|$ is bounded above), we have $\mathbb{E}[|R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f)|] \leqslant C\mathbb{E}[|\mathbf{X}_* - \mathbf{X}_{t_i}|^2]$, i.e., $R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f)$ is an order smaller than $\mathbf{X}_* - \mathbf{X}_{t_i}$. However, when f is non-globally Lipschitz (that means $|\nabla f|$ is unbounded), $R(\mathbf{X}_*, \mathbf{X}_{t_i}, \nabla f)$ may be non-negligible and require additional terms to account for its effect.

Putting the parametric flow maps in the form in (2.6), the corresponding inferred schemes (IS) with these parametrized flow maps in (2.10) are

$$\begin{aligned} &\text{IS-EM} && (\mathbf{X}_{t_{i+1}} - \mathbf{X}_{t_i})/\delta = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta f(\mathbf{X}_{t_i}) + c_2^\delta \sigma\Delta\mathbf{B}_{t_i}/\delta + \sigma_\eta \eta_i, \\ &\text{IS-RK4:} && (\mathbf{X}_{t_{i+1}} - \mathbf{X}_{t_i})/\delta = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta \phi_1^{RK4}(\mathbf{X}_{t_i}, \sigma\Delta\mathbf{B}_{t_i}) + c_2^\delta \sigma\Delta\mathbf{B}_{t_i}/\delta + \sigma_\eta \eta_i, \\ &\text{IS-SSBE} && (\mathbf{X}_{t_{i+1}} - \mathbf{X}_{t_i})/\delta = c_0^\delta \mathbf{X}_{t_i} + c_1^\delta \phi_1^{SSBE}(\mathbf{X}_{t_i}) + c_2^\delta \sigma\Delta\mathbf{B}_{t_i}/\delta + \sigma_\eta \eta_i. \end{aligned} \tag{2.13}$$

We point out that there are many other options for the parametric form. These three to-be-inferred schemes are typical: IS-EM and IS-RK4 are explicit schemes, and they will improve the statistical accuracy of the plain EM or RK4 by design (see Section 3.2). IS-RK4 is based on a multi-step scheme which provides a high-order approximation of the drift, so it tends to perform better than IS-EM when it is stable. The IS-SSBE comes from an implicit scheme, so it is likely to inherit the stability.

2.4. **Algorithm.** The following algorithm 1 summarizes the above procedure for the inference of a scheme.

3. Convergence of estimators.

We consider the convergence of the estimators in sample size in two settings: perfect model and imperfect model. The perfect model setting aims to validate our algorithm, in the sense that the algorithm can yield consistent and asymptotically normal estimators. The imperfect model setting is

Input: Full model; a high fidelity solver preserving the invariant measure.
Output: Estimated parametric scheme

1: Generate data: solve the system with the high fidelity solver, which has a small time-step dt; down sample to get time series with $\delta = \text{Gap} \times dt$. Denote the data, consisting of M independent trajectories on $[0, N\delta]$, by $\{\mathbf{X}^{(m)}_{t_0:t_N}, \mathbf{B}^{(m)}_{t_0:t_N}\}_{m=1}^{M}$ with $t_i = i\delta$.
2: Pick a parametric form approximating the flow map (2.1) as in (2.5)–(2.6).
3: Estimate parameters $c^{\delta}_{0:p}$ and σ_η as in (2.7).
4: Model selection: run the inferred scheme for cross-validation, and test the consistency of the estimators.

ALGORITHM 1. Inference-based schemes adaptive to large time-stepping (ISALT): detailed algorithm.

what we have in practice, and we show that our estimator converges to the (optimal) projection.

For simplicity of notation, we assume that $d = 1$ throughout this section. But the results also hold true entry-wisely for the system with $d > 1$.

3.1. **Convergence of estimator for perfect model.** We denote the expectation of $\bar{A}^{N,M}$ and $\bar{b}^{N,M}$ in (2.8) by A and b:

$$\begin{aligned} A &= \mathbb{E}[\bar{A}^{N,M}] = \frac{1}{N}\sum_{n=0}^{N-1}\left(\mathbb{E}\left[\langle \phi_i(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n}), \phi_j(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n})\rangle_{\mathbb{R}^d}\right]\right)_{i,j}, \\ b &= \mathbb{E}[\bar{b}^{N,M}] = \frac{1}{N}\sum_{n=0}^{N-1}(\mathbb{E}[\langle \frac{\mathbf{X}^{(m)}_{t_{n+1}} - \mathbf{X}^{(m)}_{t_n}}{\delta}, \phi_i(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n})\rangle_{\mathbb{R}^d}])_i. \end{aligned} \tag{3.1}$$

Here the expectation is with respect to the filtration generated by the initial distribution and the Brownian motion.

Assumption 3.1. *(a) Suppose that the data $\{\mathbf{X}^{(m)}_{t_0:t_N}, \mathbf{B}^{(m)}_{t_0:t_N}\}_{m=1}^{M}$ are independent trajectories of the system (2.6) with $\{\mathbf{X}^{(m)}_{t_0}\}_{m=1}^{M}$ sampled from the ergodic measure of $\mathbf{X}$. (b) Suppose that the normal matrix $\bar{A}^{N,M}$ in (2.8) and its expectation in (3.1) are invertible. (c) Suppose that the flow map $\mathcal{F}^{\delta}$ in (2.1) is square integrable.*

Theorem 3.2 (Consistency and asymptotic normality for perfect model)**.** *Under Assumption 3.1, the estimator in (2.7) converges to c^{δ} (the true parameter value) almost surely, and is asymptotically normal, when either $M \to \infty$ or $N \to \infty$:*

$$\begin{aligned} \sqrt{M}(\widehat{c^{\delta,N,M}} - c^{\delta}) &\xrightarrow{d} \mathcal{N}(0, \frac{1}{N}\sigma_\eta^2 A), \\ \sqrt{N}(\widehat{c^{\delta,N,M}} - c^{\delta}) &\xrightarrow{d} \mathcal{N}(0, \frac{1}{M}\sigma_\eta^2 A). \end{aligned} \tag{3.2}$$

Proof. By definition of $\bar{b}^{N,M}$ in (2.8) and the equation (2.6), we have

$$\begin{aligned} \bar{b}^{N,M}(i) &= \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}\langle \sum_{j=0}^{p} c_j^{\delta}\phi_j(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n}) + \sigma_\eta\eta_n^{(m)}, \phi_i(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n})\rangle_{\mathbb{R}^d} \\ &= \left(\bar{A}^{N,M}c^{\delta}\right)(i) + \bar{S}^{N,M}, \end{aligned}$$

where in the second equality we used the definition of $\bar{A}^{N,M}$ in (2.8), and we denote

$$\bar{S}^{N,M} = \frac{1}{M}\sum_{m=1}^{M} S^{N,(m)}, \text{ with } S^{N,(m)} = \frac{1}{N}\sum_{n=0}^{N-1} \langle \sigma_\eta \eta_n^{(m)}, \phi_i(\mathbf{X}_{t_n}^{(m)}, \Delta\mathbf{B}_{t_n}^{(m)})\rangle_{\mathbb{R}^d}.$$

Note that η_n is standard Gaussian and is independent of $\mathbf{B}_{t_n}$ and $\mathbf{X}_{t_n}$. Then, $S^{N,(m)}$ has mean zero and its covariance is

$$\begin{aligned}
&\mathrm{Cov}(S^{N,(m)}) \\
&\quad = \sigma_\eta^2 \frac{1}{N^2} \sum_{n,n'=0}^{N-1} \mathbb{E}\left[\langle \eta_n^{(m)}, \phi_i(\mathbf{X}_{t_n}^{(m)}, \Delta\mathbf{B}_{t_n}^{(m)})\rangle_{\mathbb{R}^d} \langle \eta_{n'}^{(m)}, \phi_i(\mathbf{X}_{t_{n'}}^{(m)}, \Delta\mathbf{B}_{t_{n'}}^{(m)})\rangle_{\mathbb{R}^d}\right] \\
&\quad = \frac{1}{N}\sigma_\eta^2 A.
\end{aligned}$$

Thus, when $M \to \infty$, we have by the Central Limit Theorem,

$$\sqrt{M}\frac{1}{M}\sum_{m=1}^{M} S^{N,(m)} \xrightarrow{d} \mathcal{N}(0, \frac{1}{N}\sigma_\eta^2 A). \tag{3.3}$$

Furthermore, $S^{N,(m)}$ is a martingale with respect to the filtration generated by $\{\mathbf{X}_{t_n}, \mathbf{B}_{t_n}, \eta_n\}$, and when $N \to \infty$, we have by martingale Central Limit Theorem [5, Theorem 3.2]

$$\sqrt{N}\frac{1}{M}\sum_{m=1}^{M} S^{N,(m)} \xrightarrow{d} \mathcal{N}(0, \frac{1}{M}\sigma_\eta^2 A). \tag{3.4}$$

We show first that, when $M \to \infty$ and for each fixed N, the estimator is consistent and asymptotically normal. Note that by the strong Law of Large Numbers, $\bar{A}^{N,M} \to A$ and $\bar{b}^{N,M} \to b$ a.s. as $M \to \infty$. Thus, $(\bar{A}^{N,M})^{-1} \to A^{-1}$ almost surely (using the fact that $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, see [29, page 22]). Then, $\widehat{c^{\delta,N,M}} = (\bar{A}^{N,M})^{-1}\bar{b}^{N,M} \to A^{-1}b$ a.s. (almost surely), i.e. the estimator is consistent. Combining (3.3) and the almost sure convergence of $(\bar{A}^{N,M})^{-1}$, we obtain the asymptotic normality by noticing that

$$\widehat{c^{\delta,N,M}} = (\bar{A}^{N,M})^{-1}\bar{b}^{N,M} = c^\delta + (\bar{A}^{N,M})^{-1}\bar{S}^{N,M}.$$

When $N \to \infty$ and M fixed, we obtain $\bar{A}^{N,M} \to A$ and $\bar{b}^{N,M} \to b$ a.s. by the ergodicity of the process. The consistency and asymptotic normality follow similarly by using (3.4). □

3.2. **Convergence of estimator for imperfect model.** In practice, the model is imperfect in our inferred scheme because we can rarely parameterize the flow map exactly. We show next that for an imperfect proposed model, the estimator converges to the projected coefficients of the flow map onto the function space spanned by the proposed basis in the ambient L^2 space. Furthermore, we show that the inferred scheme improves the statistical accuracy of the explicit scheme that it parameterizes.

Assumption 3.3. *(a) Suppose that the data* $\{\mathbf{X}_{t_0:t_N}^{(m)}, \mathbf{B}_{t_0:t_N}^{(m)}\}_{m=1}^M$ *are independent trajectories of the system* (1.1) *with* $\{\mathbf{X}_{t_0}^{(m)}\}_{m=1}^M$ *sampled from the ergodic measure of* $\mathbf{X}$. *(b) Suppose that the normal matrix* $\bar{A}^{N,M}$ *in* (2.8) *and its expectation in* (3.1) *are invertible. (c) Suppose that the flow map* $\mathcal{F}^\delta$ *in* (2.1) *is square integrable.*

The invertibility of the normal matrices $\bar{A}^{N,M}$ and A is crucial for our theory, and they lead to constraints on the basis functions. In practice, we can use it to guide the selection of basis functions and we recommend using pseudo-inverse and regularization when the normal matrix is close to singular.

With the notation A and b in (3.1), and assuming that A is invertible, we define

$$c^{\delta,\mathrm{proj}} := A^{-1}b. \tag{3.5}$$

The next lemma shows that $c^{\delta,\mathrm{proj}}$ is the projection coefficients of the flow map $\mathcal{F}^\delta$.

Lemma 3.4. *Under Assumption 3.3, the vector $c^{\delta,\mathrm{proj}}$ in (3.5) is the projection coefficients of the flow map $\mathcal{F}^\delta$ in (2.1) onto the space $\mathrm{span}\{\phi_i\}_{i=0}^p$ in $L^2(\mathbb{R}^d \times \Omega^\delta, \mu\otimes\nu)$ with μ being the invariant measure of $\mathbf{X}$ and $(\Omega^\delta, \mathcal{B}, \nu)$ being the canonical probability space for the Brownian motion $(\mathbf{B}_t, t \in [0,\delta])$.*

Proof. Note that $\mathcal{F}^\delta_{t_n} = \frac{\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n}}{\delta}$. Denote $\mathcal{F}^{\delta,m}_{t_n} = \frac{\mathbf{X}^{(m)}_{t_{n+1}} - \mathbf{X}^{(m)}_{t_n}}{\delta}$. By the definition of b in (3.1), we have

$$b(i) = \frac{1}{N}\sum_{n=0}^{N-1} \mathbb{E}\langle \mathcal{F}^{\delta,m}_{t_n}, \phi_i(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n})\rangle_{\mathbb{R}^d} = \mathbb{E}\langle \mathcal{F}^{\delta}_{t_n}, \phi_i(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})\rangle_{\mathbb{R}^d},$$

where the second equality follows from the fact that $(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})$ is stationary (so does $\mathcal{F}^\delta_{t_n}$).

Denote by $c = (c_0, c_1, \ldots, c_p)^\top$ the projection coefficients of $\mathcal{F}^\delta_{t_n}$ to $\mathrm{span}\{\phi_i\}_{i=0}^p$, and write $\mathcal{F}^\delta_{t_n} = \sum_{i=0}^p c_i\phi_i + \mathcal{F}$ with $\mathcal{F}$ satisfying $\mathbb{E}[\langle \mathcal{F}, \phi_i\rangle_{\mathbb{R}^d}] = 0$ for each $i = 0, 1, \ldots, p$. Then

$$\mathbb{E}[\langle \mathcal{F}^\delta_{t_n}, \phi_i\rangle_{\mathbb{R}^d}] = \sum_{j=0}^p c_j \mathbb{E}[\langle \phi_j, \phi_i\rangle_{\mathbb{R}^d}] = (Ac)(i).$$

Combining the above two equations, we obtain that $c^{\delta,\mathrm{proj}} = A^{-1}b = c$. □

We remark that because $\mathcal{F}^\delta(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]}, t_i, t_{i+1})$ is a functional depending on the trajectory $\mathbf{B}_{[t_i,t_{i+1}]}$, the function space of projection, $L^2(\mathbb{R}^d \times \Omega^\delta, \mu\otimes\nu)$, has an infinite dimensional state space for $(\mathbf{X}_{t_i}, \mathbf{B}_{[t_i,t_{i+1}]})$. When $\mathcal{F}^\delta_{t_n}$ depends only on $(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})$ (for instance, in the case of perfect model discussed in the previous section), the state space becomes finite dimensional and the function space is simplified to $L^2(\mathbb{R}^{2d}, \mu\otimes\nu)$ with $\nu \sim \mathcal{N}(0, \delta I_d)$.

Theorem 3.5 (Convergence of the estimator)**.** *In addition to Assumption 3.3, assume that $\mathbb{E}[|\mathcal{F}^\delta_{t_0}|^4] < \infty$ and $\mathbb{E}[|\phi_i(\mathbf{X}_{t_0}, \Delta\mathbf{B}_{t_0})|^4] < \infty$ for each $i = 0, \ldots, p$. Then, we have*

- *when $M \to \infty$ and N fixed, the estimator in (2.7) converges to the projection coefficients $c^{\delta,\mathrm{proj}}$ in (3.5) a.s. and is asymptotically normal:*

$$\sqrt{M}(\widehat{c^{\delta,N,M}} - c^{\delta,\mathrm{proj}}) \xrightarrow{d} \mathcal{N}(0, A^{-1}\Sigma^N(A^{-1})^\top), \tag{3.6}$$

where the matrix Σ^N is the covariance of

$$\widetilde{b}^{N,m}(i) = \frac{1}{N}\sum_{n=0}^{N-1} b^{n,m}, \quad \text{with } b^{n,m} = \langle \frac{\mathbf{X}^{(m)}_{t_{n+1}} - \mathbf{X}^{(m)}_{t_n}}{\delta}, \phi_i(\mathbf{X}^{(m)}_{t_n}, \Delta\mathbf{B}^{(m)}_{t_n})\rangle_{\mathbb{R}^d}.$$

- *when $N \to \infty$ and M fixed, the estimator in (2.7) also converges and is asymptotically normal*

$$\sqrt{N}(\widehat{c^{\delta,N,M}} - c^{\delta,\mathrm{proj}}) \to \mathcal{N}(0, \frac{1}{M} A^{-1}\Sigma(A^{-1})^\top), \tag{3.7}$$

with $\Sigma = \lim_{N\to\infty} N\Sigma^N$, provided that for each m,

$$\sum_{n=0}^{\infty} \mathbb{E}[b^{n,m}\mathbb{E}[b^{k,m}|\mathcal{M}_0]] \text{ converges for each } k \geqslant 0 \text{ and}$$

$$\lim_{N\to\infty} \sum_{k=K}^{\infty} \mathbb{E}[b^{k,m}\mathbb{E}[b^{N,m}|\mathcal{M}_0]] = 0 \text{ uniformly in } K,$$

where $\mathcal{M}_0$ denotes the filtration generated by the extended stationary process up to time t_0.

Proof. When $M \to \infty$, by the strong Law of Large Numbers, we have $\bar{A}^{N,M} \to A$ and $\bar{b}^{N,M} \to b$ a.s. when $M \to \infty$. Thus, $\widehat{c^{\delta,N,M}} = (\bar{A}^{N,M})^{-1}\bar{b}^{N,M} \to A^{-1}b = c^{\delta,\mathrm{proj}}$ a.s. according to Lemma 3.4. To prove the asymptotic normality, note that for each m, the random vector $\widetilde{b}^{N,m}$ with entries

$$\widetilde{b}^{N,m}(i) = \frac{1}{N}\sum_{n=0}^{N-1} \langle \mathcal{F}_{t_n}^{\delta,m}, \phi_i(\mathbf{X}_{t_n}^{(m)}, \Delta\mathbf{B}_{t_n}^{(m)})\rangle_{\mathbb{R}^d}$$

has mean $\mathbb{E}[\widetilde{b}^{N,m}] = b$ and covariance Σ^N with entries $\Sigma^N_{i,j} = \mathbb{E}[\widetilde{b}^{N,m}(i)\widetilde{b}^{N,m}(j)] - b(i)b(j)$. Here the covariance exists because

$$\begin{aligned}\mathbb{E}[\widetilde{b}^{N,m}(i)\widetilde{b}^{N,m}(j)] &\leqslant \max_i \mathbb{E}[|\widetilde{b}^{N,m}(i)|^2] \leqslant \max_i \mathbb{E}[|\langle \mathcal{F}_{t_n}^{\delta,m}, \phi_i(\mathbf{X}_{t_n}^{(m)}, \Delta\mathbf{B}_{t_n}^{(m)})\rangle_{\mathbb{R}^d}|^2] \\ &\leqslant (\mathbb{E}[|\mathcal{F}_{t_0}^{\delta}|^4])^{1/2} \max_i (\mathbb{E}[|\phi_i(\mathbf{X}_{t_0}, \Delta\mathbf{B}_{t_0})|^4)^{1/2}.\end{aligned}$$

Then, $\bar{b}^{N,M}$ is the average of M iid samples $\{\widetilde{b}^{N,m}\}$, each of which has covariance Σ^N. Hence, by the Central Limit Theorem, we have

$$\sqrt{M}(\bar{b}^{N,M} - b) \xrightarrow{d} \mathcal{N}(0, \Sigma^N).$$

Combining with the fact that $\bar{A}^{N,M} \to A$ a.s. and that these matrices are invertible, we obtain (3.6).

When $N \to \infty$, we obtain the convergence and asymptotic normality by ergodicity. First, by ergodicity, we have $\bar{A}^{N,M} \to A$ and $\bar{b}^{N,M} \to b$ a.s. as $M \to \infty$. Thus, $\widehat{c^{\delta,N,M}} = (\bar{A}^{N,M})^{-1}\bar{b}^{N,M} \to A^{-1}b = c^{\delta,\mathrm{proj}}$ a.s. (almost surely). Next, to prove the asymptotic normality, note that by the Central Limit Theorem for stationary processes [8, Theorem 1], we have $\Sigma = \lim_{N\to\infty} N\Sigma^N$ and

$$\sqrt{N}(\bar{b}^{N,M} - b) \xrightarrow{d} \mathcal{N}(0, \frac{1}{M}\Sigma).$$

Then we obtain (3.7) by noting that $\bar{A}^{N,M} \to A$ a.s. as above. □

We show next that when the inferred scheme is a parametrization of an explicit scheme, it leads to improvements in the sense that the inferred schemes's 1-step numerical error, which depends on the residual of regression, is bounded above by the explicit scheme's.

Theorem 3.6 (Bounds for residual)**.** *Assume that an inferred scheme* (2.6) *parameterizes an explicit scheme, e.g., the* IS-EM *or* IS-RK4 *in* (2.13) *from its explicit scheme in* (2.3)*. Then, with Assumption* 3.3*, we have*

- *(Bounds for residual) the regression residual in* (2.7) *satisfies*

$$\mathbb{E}(\widehat{\sigma^{N,M}})^2 \leqslant \frac{2}{d}\mathbb{E}\Big[|\frac{\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}}{\delta} - F^\delta(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2\Big], \tag{3.8}$$

  where $F^\delta(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})$ *denotes the flow map of the explicit scheme, such as* F^δ_{EM} *or* F^δ_{RK4} *in* (2.3)*.*
- *(convergence of residual deviation) Furthermore, assuming the conditions in Theorem* 3.5*, we have*

$$(\widehat{\sigma^{N,M}})^2 \to \frac{2}{d\delta^2}\mathbb{E}|\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}-\delta F^\delta(c^{\delta,\mathrm{proj}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2 \quad a.s.$$

  if either $N\to\infty$ *or* $M\to\infty$*, for each* δ*.*

Proof. We write the flow map of the explicit scheme in a parametric form, that is, $F^\delta(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n}) = F^\delta(c^*,\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})$ as in (2.5). Then, since the estimator $\widehat{c^{\delta,N,M}}$ in (2.7) is the minimizer of the likelihood, we have

$$\begin{aligned}(\widehat{\sigma^{N,M}})^2 &= \frac{2}{d\delta^2}\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}|\mathbf{X}^{(m)}_{t_{n+1}}-\mathbf{X}^{(m)}_{t_n}-\delta F^\delta(\widehat{c^{\delta,N,M}},\mathbf{X}^{(m)}_{t_n},\Delta\mathbf{B}^{(m)}_{t_n})|^2\\ &\leqslant \frac{2}{d\delta^2}\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}|\mathbf{X}^{(m)}_{t_{n+1}}-\mathbf{X}^{(m)}_{t_n}-\delta F^\delta(c^*,\mathbf{X}^{(m)}_{t_n},\Delta\mathbf{B}^{(m)}_{t_n})|^2.\end{aligned} \tag{3.9}$$

Since the process $(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})_n$ is stationary, we have

$$\begin{aligned}\mathbb{E}(\widehat{\sigma^{N,M}})^2 &= \frac{2}{d\delta^2}\mathbb{E}|\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}-\delta F^\delta(\widehat{c^{\delta,N,M}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2\\ &\leqslant \frac{2}{d}\mathbb{E}|\frac{\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}}{\delta}-F^\delta(c^*,\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2.\end{aligned} \tag{3.10}$$

Recalling that $F^\delta(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n}) = F^\delta(c^*,\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})$, we have (3.8).

Next, consider the convergence of $(\widehat{\sigma^{N,M}})^2$. To simplify the notations, let

$$\begin{aligned}&\Delta\mathbf{X}^{(m)}_{t_n} = \mathbf{X}^{(m)}_{t_{n+1}}-\mathbf{X}^{(m)}_{t_n}, \quad F^{(m)}(c) = F^\delta(c,\mathbf{X}^{(m)}_{t_n},\Delta\mathbf{B}^{(m)}_{t_n}),\\ &\Delta F^{(m)} = F^{(m)}(\widehat{c^{\delta,N,M}}) - F^{(m)}(c^{\delta,\mathrm{proj}}).\end{aligned}$$

Note that as either $N\to\infty$ or $M\to\infty$, we have $\Delta F^{(m)}\to 0$ a.s. for each m according to Theorem 3.5; hence,

$$\begin{aligned}&|\Delta\mathbf{X}^{(m)}_{t_n}-\delta F^{(m)}(\widehat{c^{\delta,N,M}})|^2 = |\Delta\mathbf{X}^{(m)}_{t_n}-\delta\left(F^{(m)}(c^{\delta,\mathrm{proj}})+\Delta F^{(m)}\right)|^2\\ =&|\Delta\mathbf{X}^{(m)}_{t_n}-\delta F^{(m)}(c^{\delta,\mathrm{proj}})|^2\\ &\quad -2\delta(\Delta\mathbf{X}^{(m)}_{t_n}-\delta F^{(m)}(c^{\delta,\mathrm{proj}}))\Delta F^{(m)}(c^{\delta,\mathrm{proj}})+\delta^2|\Delta F^{(m)}|^2\\ \to&|\Delta\mathbf{X}^{(m)}_{t_n}-\delta F^{(m)}(c^{\delta,\mathrm{proj}})|^2 \quad a.s.\end{aligned}$$

Then, from the equation in (3.9), we have,

$$\begin{aligned}(\widehat{\sigma^{N,M}})^2 =& \frac{2}{d\delta^2}\frac{1}{MN}\sum_{m=1}^{M}\sum_{n=0}^{N-1}|\Delta \mathbf{X}_{t_n}^{(m)} - \delta F^{(m)}(\widehat{c^{\delta,N,M}})|^2 \\ \to& \frac{2}{d}\mathbb{E}\left|\frac{\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}}{\delta} - F^{\delta}(c^{\delta,\mathrm{proj}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})\right|^2 \quad a.s.\end{aligned}$$

by the Law of Large Numbers. This completes the proof. □

Remark 3.7 (Order of residuals for IS-RK4 and IS-EM)**.** As $\delta \to 0$, the residual for an inferred scheme converges to zero at the order of the explicit scheme that it parameterizes. Recall that either Euler-Maruyama scheme or the HRK4 schemes has $\mathbb{E}\big[|\frac{\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}}{\delta} - F^{\delta}(c^*,\mathbf{X}_{t_i},\Delta\mathbf{B}_{t_n})|^2\big] = O(\delta)$, which follows from the Itô formula. Thus, for the inferred schemes IS-EM and IS-RK4 in (2.3), we have $\mathbb{E}[(\widehat{\sigma^{N,M}})^2] = O(\delta)$. Furthermore, by the strong Law of Large Numbers, we have $(\widehat{\sigma^{N,M}})^2 \to \mathbb{E}(\widehat{\sigma^{N,M}})^2$ a.s. when either $M \to \infty$ or $N \to \infty$. Thus, the estimator $\widehat{\sigma^{N,M}} = O(\delta^{1/2})$ a.s. for large N or M. However, IS-SSBE's residual may not converge to zero, because SSBE is not in the parametric family of IS-SSBE and the neglected term R in (2.12) can prevent the residual from decaying to zero (see Figure 5(b)).

Corollary 3.8 (Error of the inferred scheme)**.** *The inferred scheme has the same strong order as the explicit scheme that it parameterizes. More precisely, consider the inferred scheme with parameters* $\widehat{c^{\delta,N,M}}$ *and* $\widehat{\sigma^{N,M}}$*:*

$$\mathbf{Y}_{n+1} - \mathbf{Y}_n = \delta F^{\delta}(\widehat{c^{\delta,N,M}},\mathbf{Y}_n,\Delta\mathbf{B}_{t_n}) + \delta\widehat{\sigma^{N,M}}\eta_n,$$

where η_n *is* $\mathcal{N}(0,I_d)$ *and is independent of* $\{\mathbf{B}_{t_n}\}_{n\geqslant 1}$*. Assume that it is a parametrization of an explicit scheme with flow map* $F^{\delta}(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})$*. Suppose that the parameters satisfy* $|\widehat{c^{\delta,N,M}} - c^{\delta,\mathrm{proj}}| \leqslant \epsilon$ *and*

$$|(\widehat{\sigma^{N,M}})^2 - \frac{2}{d\delta^2}\mathbb{E}|\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^{\delta}(c^{\delta,\mathrm{proj}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2 \leqslant \epsilon^2$$

with $\epsilon \in (0,1)$ *(which exists by Theorem 3.5–3.6). Then, with* $\mathbf{Y}_n = \mathbf{X}_{t_n}$*, the 1-step error satisfies*

$$\mathbb{E}|\mathbf{X}_{t_{n+1}} - \mathbf{Y}_{n+1}|^2 \lesssim \frac{2}{d}\mathbb{E}\Big[|\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^{\delta}(\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2\Big] + \delta^2\epsilon.$$

In other words, the inferred scheme's 1-step error is smaller than the explicit scheme's, provided that the error in the estimator of parameters is negligible.

Proof. With $\mathbf{Y}_n = \mathbf{X}_{t_n}$, the 1-step error is

$$\mathbf{X}_{t_{n+1}} - \mathbf{Y}_{n+1} = \mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^{\delta}(\widehat{c^{\delta,N,M}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n}) - \delta\widehat{\sigma^{N,M}}\eta_n.$$

Thus, taking expectation (conditional on the parameters given), we have

$$\mathbb{E}|\mathbf{X}_{t_{n+1}} - \mathbf{Y}_{n+1}|^2 \leqslant \mathbb{E}|\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^{\delta}(\widehat{c^{\delta,N,M}},\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n})|^2 + \delta^2\widehat{\sigma^{N,M}}^2.$$

Note that we cannot apply (3.10) to bound the first term, because the above expectation is conditional on the parameters, whereas (3.10) is not. To control the first term, again denoting

$$\Delta\mathbf{X}_{t_n} = \mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n}, \quad F(c) = F^{\delta}(c,\mathbf{X}_{t_n},\Delta\mathbf{B}_{t_n}), \quad \Delta F = F(\widehat{c^{\delta,N,M}}) - F(c^{\delta,\mathrm{proj}}),$$

we can write $|\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^\delta(\widehat{c^{\delta,N,M}}, \mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})|^2$ as

$$\begin{aligned}|\Delta\mathbf{X}_{t_n} - \delta F(\widehat{c^{\delta,N,M}})|^2 &= |\Delta\mathbf{X}_{t_n} - \delta F(c^{\delta,\mathrm{proj}}) - \delta\Delta F|^2\\ &\leqslant 2|\Delta\mathbf{X}_{t_n} - \delta F(c^{\delta,\mathrm{proj}})|^2 + 2\delta^2|\Delta F|^2.\end{aligned}$$

Note that $\mathbb{E}|\Delta F|^2 \lesssim |\widehat{c^{\delta,N,M}} - c^{\delta,\mathrm{proj}}|^2 \leqslant \epsilon^2$. Thus, we have

$$\mathbb{E}|\mathbf{X}_{t_{n+1}} - \mathbf{Y}_{n+1}|^2 \lesssim \quad \mathbb{E}|\Delta\mathbf{X}_{t_n} - \delta F(c^{\delta,\mathrm{proj}})|^2 + \delta^2\epsilon^2.$$

Meanwhile, note that by combining the two items of Theorem 3.6, we have

$$\mathbb{E}|\Delta\mathbf{X}_{t_n} - \delta F(c^{\delta,\mathrm{proj}})|^2 \leqslant \frac{2}{d}\mathbb{E}\Big[|\mathbf{X}_{t_{n+1}} - \mathbf{X}_{t_n} - \delta F^\delta(\mathbf{X}_{t_n}, \Delta\mathbf{B}_{t_n})|^2\Big].$$

This completes the proof. □

Remark 3.9. The idea of inference-based scheme also applies to non-ergodic systems to obtain reduced-in-time models. The convergence of the parameters in Theorem 3.5 and Theorem 3.6 remains true when the sample size M goes to infinity. Furthermore, one may accelerate the simulation of slowly converging ergodic systems by training inference-based schemes iteratively in time. In this study, we focus on non-globally Lipschitz ergodic systems to highlight the ability of the inferred-scheme in reproducing long-term statistics.

Remark 3.10 (Optimal reduction in time)**.** We emphasize that our goal is to infer an explicit scheme with a relative large time-step for efficient simulation of non-globally Lipschitz ergodic systems. Corollary 3.8 indicates that the inferred scheme's 1-step error (i.e., the approximate error of the flow map $\mathcal{F}^\delta$ in (2.1)) decays with improved speed as the time-step decreases. But a smaller 1-step error due to a smaller time-step does not necessarily imply a better inferred scheme, because the 1-step error from the residual can be accumulated in more iterations (e.g., the EM scheme can have a wrong invariant measure). To improve the inferred scheme, we seek an optimal time-step that balances the 1-step error and the accumulated residual into the invariant measure. In our examples in the next section, the inferred scheme performs the best (at reproducing the invariant measure and temporal correlation) when the time-step is moderately large. This is similar to the parameter estimation for homogenization of multiscale process [35], where the sub-sampling rate must be between the two characteristic time scales of the SDEs.

4. Examples.

In this section, we test and compare three benchmark examples for each inference-based scheme proposed in (2.13) using two different parametric settings: c_0 excluded vs c_0 included, so as to distinct the contribution of the linear term parameterized by c_0. Three non-globally Lipschitz examples are: a 1D system with the double-well potential; a 2D gradient system; and a 3D stochastic Lorenz system with degenerate noises.

In each of the examples, we generate data for inference by the Split Step Backward Euler (SSBE) scheme with a fine time-step Δt. We infer schemes for different time step-sizes $\delta = \mathrm{Gap} \times \Delta t$ with 10 options for the time gap: $\mathrm{Gap} \in \{1, 2, 4, 10, 20, 40, 80, 120, 160, 200\}$, which will be used to select optimal time gap and demonstrate the convergence order of the residual in Theorem 3.6. The computations of inference include 5 options: (1) IS-EM with c_0 excluded; (2) IS-RK4 with c_0 excluded; (3) IS-RK4 with c_0 included; (4) IS-SSBE with c_0 excluded; and (5) IS-SSBE with c_0 included.

We assess the performance of these schemes by the accuracy of the reproduced invariant density (PDF) and the auto-correlations function (ACF), which are empirically computed from a long trajectory. The accuracy of PDF is measured by the total variation distance (TVD) from the reference PDF of data.

Once we identify the best performing scheme for each example, we fix the inference settings and present the convergence of the estimators and the residuals with respect to the time Gap as well as the number of trajectories M.

In summary, we find from the examples that

- The inferred scheme has significantly stronger numerical stability than the plain schemes. The IS-RK4 and IS-SSBE exhibit better stability than IS-EM. In particular, they can tolerate time-steps that are significantly larger than the plain RK4 or SSBE. Specifically, we find the plain RK4 and SSBE always blow up even when Gap = 20, whereas the inferred schemes are still stable when Gap is larger than 200, which improves the efficiency by an order of more than 10. We summarize the time gaps of blow-up for plain verse inferred schemes for each example in the following table.

	1D double-well	2D gradient system	3D Lorenz system
Plain RK4	Gap = 20	Gap = 20	Gap = 10
IS-RK4	Gap > 200	Gap > 200	Gap > 400
Plain SSBE	Gap = 40	Gap = 40	Gap = 20
IS-SSBE	Gap > 200	Gap > 200	Gap > 400

TABLE 2. Time gap of blow-up for each scheme: plain verse inferred.

- The inferred scheme can reproduce the invariant measure accurately. Both IS-RK4 and IS-SSBE perform well when the stochastic force dominates the dynamics. But when the drift dominates the dynamics in the example of Lorenz system, IS-RK4 performs better than IS-SSBE, because it provides a better approximation to the drift than IS-SSBE.
- The inferred scheme reproduces the invariant density the best when the time-step is medium large (with a time gap between Gap = 80 and Gap = 160), suggesting a balance between the approximation error of the flow map and the numerical error in simulating the invariant density. It is open to have an *a-priori* estimate of the optimal time gap.

4.1. **1D double-well potential.** First consider a 1D SDE with a double-well potential [34]

$$dX_t = -V'(X_t)dt + \sqrt{2/\beta}dB_t, \tag{4.1}$$

with $V(x) = \frac{\mu}{4}(x^2-1)^2$. The corresponding invariant measure is $\frac{1}{Z}\exp^{-\beta V(x)}$ where Z being the normalizing constant $Z := \int_{\mathbb{R}} \exp^{-\beta V(x)}\, dx$. We set $\mu = 2$ and $\beta = 1$.

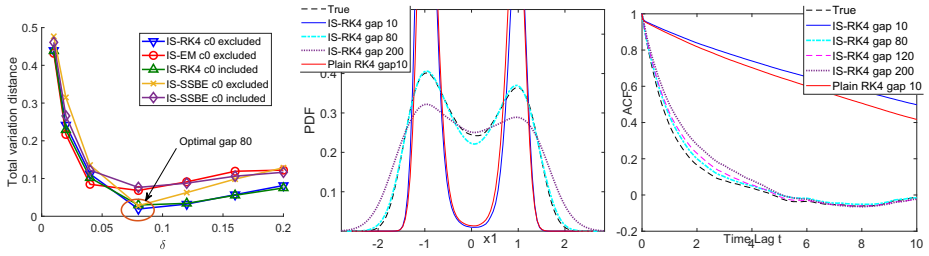
We generate data by SSBE with a fine time-step $\Delta t = 1e-3$. We first simulate a long trajectory on an interval $[0, T]$ with $X_0 = 1/2$ and $T = 2000$ (i.e., two million time steps), which is found to be long enough to represent the invariant density (PDF). This long trajectory will also provide us the reference PDF and ACF, which are referred as the true values to be approximated. Then we generate $M = 1000$ trajectories on the time interval $[0, 40]$ with initial conditions sampled

from the long trajectory. The data are the M trajectories of the Brownian motion and the process (X_t) observed at discrete times $\{t_n = n\delta = n\text{Gap} \times \Delta t\}$, as in (2.4).

The parameters of the schemes in (2.13) are then estimated by Algorithm 1 for each $\delta = \text{Gap} \times \Delta t$.

Figure 2(a) shows the TVD of the five inferred schemes with coarse verse fine Gap $\in \{10, 20, 40, 80, 120, 160, 200\}$. Note that for every scheme, the TVD first decreases and then increases, reaching the smallest TVD when Gap $= 80$. This suggests that when the gap is small, the approximation error of the flow map (recall that the data are from an implicit scheme while the inferred schemes are explicit schemes) dominates the error in the invariant measure; when the gap is large, the numerical error of the inferred schemes dominates the TVD. A balance between the two errors is reached at the medium large time-step.

We first select the scheme that reproduces the invariant density with the smallest TVD. Overall, the IS-RK4 schemes perform the best and the inclusion of c_0 brings in negligible improvement. Thus, we select IS-RK4 without c_0 to demonstrate further results.

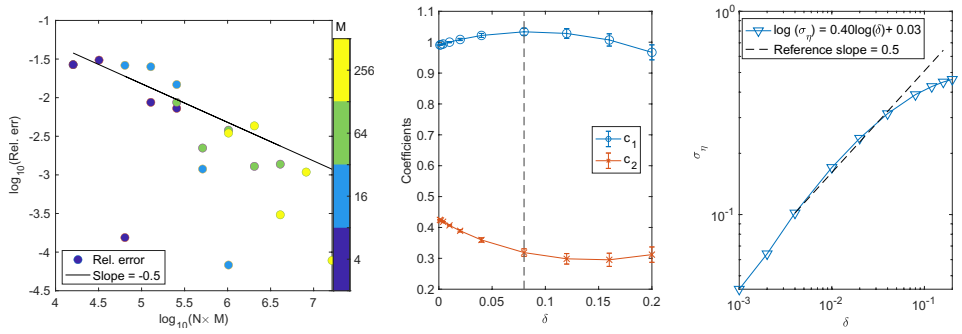


(a) TVD (b) PDF (c) ACF

FIGURE 2. Large-time statistics for 1D double-well potential. (a) TVD between the empirical invariant densities (PDF) of the inferred schemes and the reference PDF from data. (b) and (c): PDFs and ACFs comparison between the IS-RK4 with c_0 excluded and the reference data.

Figure 2 (b-c) show the PDFs and auto-correlation functions (ACFs) of IS-RK4 with c_0 excluded at three representative time gaps Gap $\in \{10, 80, 200\}$, in comparison with those of the reference data and the plain RK4 with Gap $= 10$. When Gap is small, that is Gap $= 10$, the IS-RK4 is close to the plain RK4, and both produce PDFs and ACFs with large errors. The PDF and ACF generated by IS-RK4 with Gap $= 80$ is the best among all used gaps, fitting the true PDF and ACF almost perfectly. Furthermore, when time gap is as large as Gap $= 200$, the IS-RK4 can still produce qualitative results with the feature of PDF (that is the double-well feature), whereas the plain RK4 scheme blows up when Gap $= 20$.

We also test the convergence of the estimators in sample size and their dependence on the time-step, as well as the order of residual, aiming to confirm the theory in Section 3. Figure 3(a) shows that the relative error of $\widehat{c_1^{\delta,N,M}}$ converges at a rate about $(MN)^{-1/2}$ as the sample size N or M increases. Here we take the estimator from the largest sample size as the projection coefficient, and compute the relative error to it. Note that the estimator of c_1 is close to 1. Thus, the estimator $\widehat{c_1^{\delta,N,M}}$ converges at a rate about $(MN)^{-1/2}$, matching Theorem 3.5. The convergence of the estimator of c_2 has similar convergence rate.



(a) Convergence of c_1 in sample size (b) Coefficients and residuals

FIGURE 3. 1D double-well potential: Convergence of estimators in IS-RK4 with c_0 excluded. (a) The relative error of the estimator $\widehat{c_1^{\delta,N,M}}$ with $\delta = 80 \times \Delta t$ converges at an order about $(MN)^{-1/2}$, matching Theorem 3.5. (b) Left column: The coefficients depend on the time-step $\delta = \text{Gap} \times \Delta t$, with c_1 being almost 1 and c_2 being close to linear in δ until $\delta > 0.08$. The error bars, which are too narrow to be seen, are the standard deviations of the single-trajectory estimators from the M-trajectory estimator. Right column: The residual decays at an order $O(\delta^{1/2})$, matching Theorem 3.6.

Figure 3(b) shows the dependence of the estimators on the time-step $\delta = \text{Gap} \times \Delta t$. The coefficient c_1 is almost 1, while c_2 is close to being linear in δ. Furthermore, it also shows that the estimators from each single trajectory are close to the M-trajectory estimator, with small standard deviations represented by error bars that are too narrow to be seen. The residual decays at an order about 0.49 with respect to δ, closely matching the rate stated in Theorem 3.6.

4.2. **A 2D gradient system.** Consider a 2D dissipative gradient system [34]

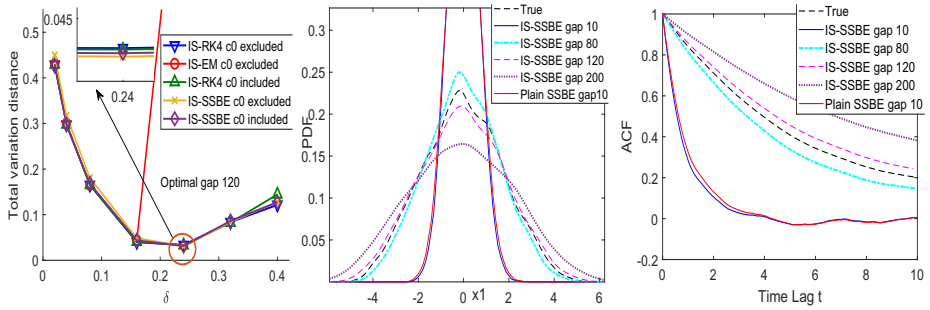
$$d\mathbf{X}_t = -\nabla V(\mathbf{X}_t)dt + \sqrt{2/\beta}d\mathbf{B}_t, \tag{4.2}$$

with $V(\mathbf{X}) = V(x_1, x_2) = \exp\left(\frac{\mu_1}{2}x_1^2 + \frac{\mu_2}{2}x_2^2\right)$. The corresponding invariant measure is $\frac{1}{Z}exp^{-\beta V(x_1,x_2)}$ where Z being the normalizing constant that is computed by $Z := \int_{\mathbb{R}^2} exp^{-\beta V(x_1,x_2)}dx_1dx_2$. We set $\mu_1 = 0.1$, $\mu_2 = 1$ and $\beta = 2$. Because $\mu_2 = 10\mu_1$, so x_1 is a slowly evolving variable compared to x_2 and the resulting dynamics displays a multi-scale feature. Consequently, we estimate parameters entry-wisely and we focus on the marginal invariant density of x_1.

We generate data by the SSBE scheme with $\Delta t = 2e-3$ and time interval $[0, 2000]$ with total time steps $tN = 1e6$. The rest setting and procedure are the same as the 1D double-well potential case.

Figure 4(a) shows that IS-RK4 and IS-SSBE schemes have comparable TVD, and they reach the minimal TVD when Gap = 120, where IS-EM blows up. They produce similar PDFs and ACFs, so we only present those of IS-SSBE with c_0 excluded. Figure 4(b-c) show the PDFs and ACFs at representative time gaps Gap $\in \{10, 80, 120, 200\}$. The findings are similar to those for the 1D double-well potential: (i) the performance of IS-SSBE first improves and then deteriorates as Gap increases; (ii) IS-SSBE can tolerate significantly larger time-step than the plain SSBE, where the plain SSBE blows up due to the Newton-Raphson method used as

the implicit solver, which can only tolerate a small time-step limited by the inversion (similar to (2.11)) in the Newton-Raphson method in the implicit solver.



(a) TVD (b) PDF (c) ACF

FIGURE 4. Large-time statistics for the 2D gradient system. (a) TVD between the x_1 marginal invariant densities (PDF) of the inferred schemes and the reference PDF from data. (b) and (c): PDFs and ACFs comparison between IS-SSBE with c_0 excluded and the reference data.

The convergence of the estimators in sample size is roughly of order $(MN)^{-1/2}$, as shown in Figure 5(a). Figure 5(b) shows that the estimators of c_1 and c_2 depend almost linearly on δ. Also, c_1's single-trajectory estimators have negligible standard deviations from the M-trajectory estimator, while c_2's estimators have a persistent noticeable standard deviation. This suggests that IS-SSBE has large uncertainties in the stochastic force term (recall that c_1 and c_2 being the coefficients of the scaled drift and the stochastic force, see (2.13)). In the right column, the residual of IS-SSBE remains little changed when δ decreases, far from a decay rate 0.5. This does not violate Theorem 3.6, which is for parametrizations of explicit schemes. Instead, this highlights that the IS-SSBE is not a parametrization of the SSBE implicit scheme, and it has a flow map $\widetilde{F}^{\delta}_{SSBE}$ with distance to the true flow map $\mathbb{E}\big[|\frac{\mathbf{X}_{t_{n+1}}-\mathbf{X}_{t_n}}{\delta} - \widetilde{F}^{\delta}_{SSBE}(c, \mathbf{X}_{t_i}, \Delta\mathbf{B}_{t_n})|^2\big]$ depending little on δ. Such a feature can be helpful for further improving the parametric form.

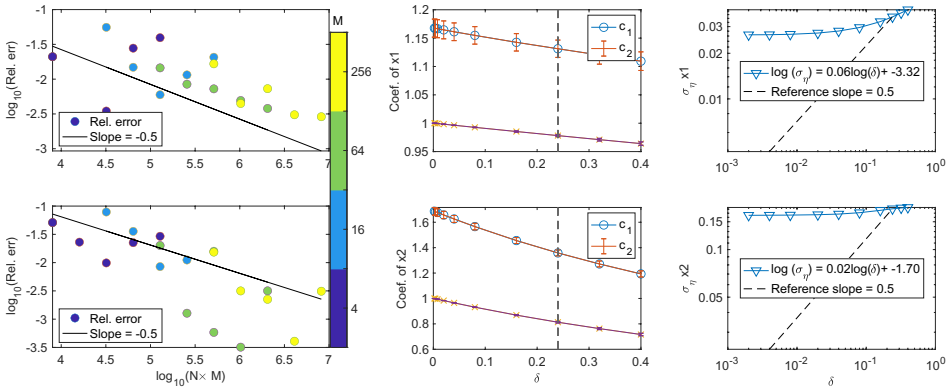
Figure 6 shows the convergence of the estimator for IS-RK4. Similar to the 1D case, we observe a convergence rate $(MN)^{-1/2}$ in Figure 6(a). Also, in Figure 6(b), we observe almost δ independent estimators and the expected decay rate $O(\delta^{1/2})$ of residuals proved in Theorem 3.6.

4.3. **Stochastic Lorenz system with degenerate noise.** Consider next the 3D stochastic Lorenz system with degenerate noise [34]

$$\begin{aligned} dx_1 &= \sigma(x_2 - x_1)dt + \sqrt{2/\beta}dB_1, \\ dx_2 &= \big(x_1(\gamma - x_3) - x_2\big)dt + \sqrt{2/\beta}dB_2, \\ dx_3 &= (x_1x_2 - bx_3)dt. \end{aligned} \tag{4.3}$$

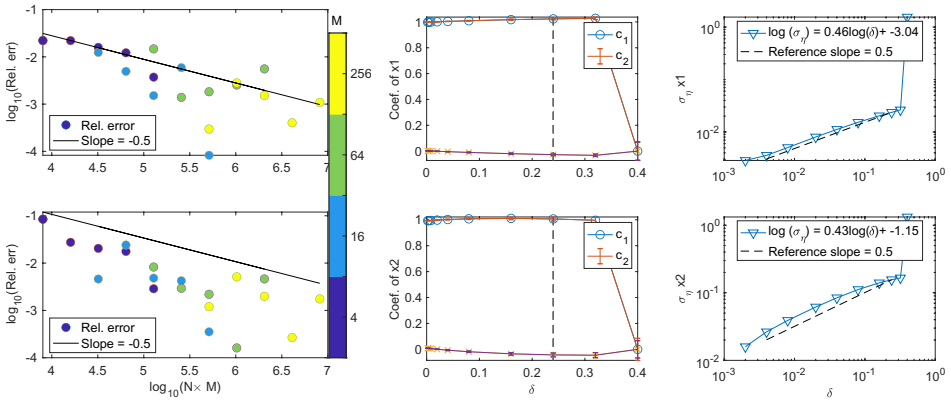
We set $\sigma = 10$, $\gamma = 28$, $b = 8/3$ and $\beta = 1$. This stochastic chaotic system is exponentially ergodic with a regular invariant measure because it is dissipative and hypoelliptic.

As before, we generate data by SSBE with $\Delta t = 5e-4$ and a reference long trajectory with $tN = 6e6$ time steps (or equivalently, on the time interval $[0, 3000]$).



(a) Convergence of c_1 in sample size (b) Coefficients and residuals

FIGURE 5. 2D gradient system: Convergence of estimators in IS-SSBE with c_0 excluded. (a) The relative error of the estimator $\widehat{c_1^{\delta,N,M}}$ with $\delta = 120\Delta t$ converges at an order about $(MN)^{-1/2}$, matching Theorem 3.5. (b) Left column: The estimators of c_1, c_2 are almost linear in δ. Right column: The residual changes little as δ decreases, due to that IS-SSBE is not a parametrization of an explicit scheme (thus, Theorem 3.6 does not apply).

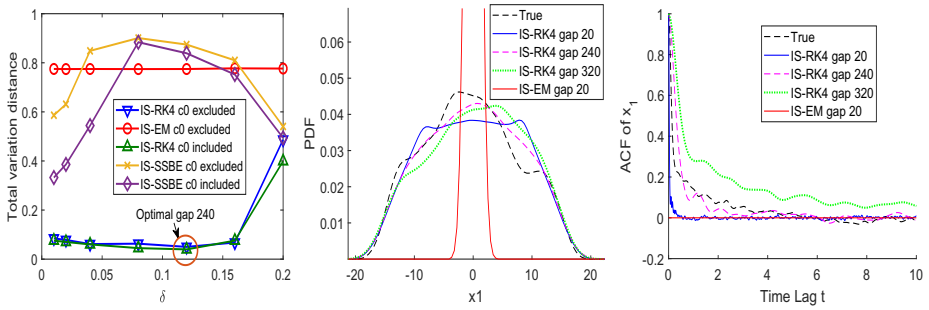


(a) Convergence of c_1 in sample size (b) Coefficients and residuals

FIGURE 6. 2D gradient system: Convergence of estimators in IS-RK4 with c_0 excluded. (a) The relative error of the estimator $\widehat{c_1^{\delta,N,M}}$ with $\delta = 120\Delta t$ converges at an order about $(MN)^{-1/2}$, matching Theorem 3.5. (b) Left column: The estimators of c_1, c_2 are constant for all δ. Right column: The residual decays at an order $O(\delta^{1/2})$, matching Theorem 3.6.

We consider time gaps Gap $\in \{20, 40, 80, 160, 240, 320, 400\}$, so the maximal time-step is still 0.2.

Figure 7(a) shows the TVD of the inferred schemes. This time, the IS-RK4 scheme performs significantly better than IS-SSBE schemes, with relatively small TVD for most time gaps. This is due to the high-order approximation of RK4 to

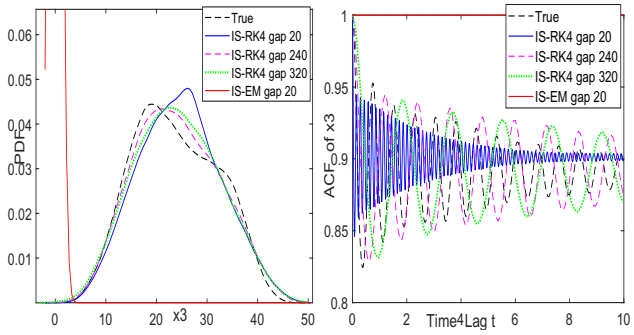


(a) TVD (b) PDF (c) ACF

FIGURE 7. Large-time statistics of x_1 for the stochastic Lorenz system. (a) TVD between the x_1 marginal invariant densities (PDF) of the inferred schemes and the reference PDF from data. (b) and (c): PDFs and ACFs comparison between IS-RK4 with c_0 included and the reference data.

the drift, particularly when the drift dominates the dynamics (note that the state variable x_1 is at a scale of magnitude larger than the degenerate noise). The IS-RK4 with c_0 included performs the best and we select it for further demonstration of results.

Figure 7(b-c) show the PDFs and ACFs at representative time gaps Gap $\in \{20, 240, 320\}$. Since the plain RK4 blows up at Gap $= 10$, so we display the results from IS-EM instead. The findings are similar to those for the 1D double-well potential: (i) the performance of IS-RK4 first improves and then deteriorates as Gap increases; (ii) IS-RK4 can tolerate significantly larger time-step than the plain RK4.



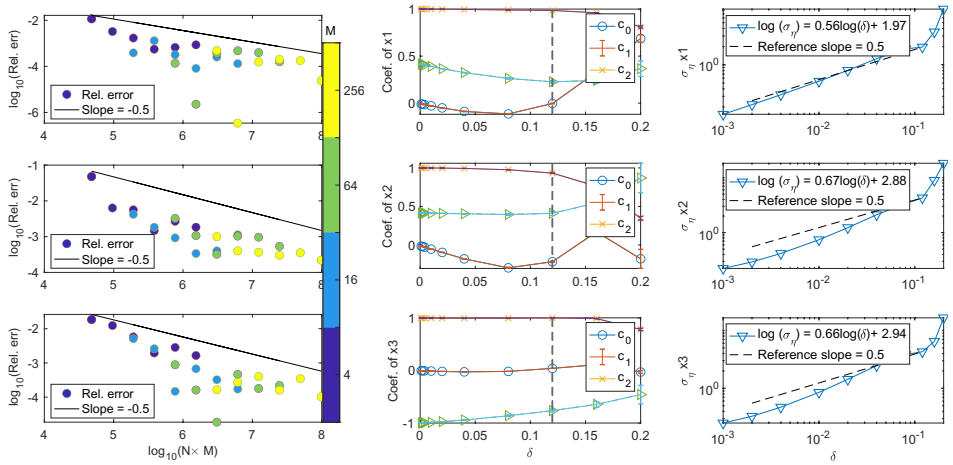
(a) PDF (b) ACF

FIGURE 8. ACF and PDF of x_3 in the stochastic Lorenz system. Similar to the other examples, IS-RK4 (with c_0 included) reproduces the PDF and the ACF the best when the time-step is medium large, while plain RK4 and IS-EM blow up even when Gap = 20.

Moreover, we also plot the PDF and ACF of x_3 in Figure 8. The dynamics of x_3 is the most challenging because there is no diffusive stochastic force acting on it and its ACF is highly oscillatory. As usual, the IS-RK4 can reproduce the PDF and ACF well, whereas the plain RK4 and IS-EM blow up even when the time-step is small. In particular, the IS-RK4 produces the periodic and decay feature of x_3's

ACF when the time-step is medium large, that is Gap = 240. We expect the best performance to be achieved at a gap between 120 to 240, and we plan to further study the optimal time gap and other improvements.

The IS-RK4 has convergence results mostly as expected. Figure 9(a) shows that the estimators of c_1 for each entry of (x_1, x_2, x_3) converge at an almost perfect rate $(NM)^{-1/2}$. Figure 9(b) shows that the estimator of c_0, c_1, c_2 remains little varied until $\delta = 0.12$ (i.e., Gap > 240) for each entry. It also shows that the residuals of all three entries decay at a rate slightly higher than $O(\delta^{1/2})$.



(a) Convergence of c_1 in sample size (b) Coefficients and residuals

FIGURE 9. The 3D stochastic Lorenz system: Convergence of estimators in IS-RK4 with c_0 included. (a) The relative error of the estimator $\widehat{c_1^{\delta,N,M}}$ with $\delta = 240\Delta t = 0.12$ converges at order about $(MN)^{-1/2}$, matching Theorem 3.5. (b) Left column: The estimators of c_0, c_1, c_2 are varies little until $\delta > 0.12$. The vertical dash line is the optimal time gap. Right column: The residuals decay at orders slightly higher than $O(\delta^{1/2})$.

5. **Conclusions and outlook.** We have introduced a general framework to infer schemes adaptive to large time-stepping (ISALT) from data for locally Lipschitz ergodic SDEs. We formulate it as a statistical learning problem, in which we learn an approximation to the infinite-dimensional discrete-time flow map. By deriving informed basis functions from classical numerical schemes, we obtain a low-dimensional parameter estimation problem, avoiding the curse of dimensionality in statistical learning.

Under mild conditions, we show that the estimator converges as the data size increases, and the inferred scheme has the same 1-step strong order as the explicit scheme it parameterizes. Thus, the inferred scheme comes with improved performance guaranteed. Numerical tests on three non-globally Lipschitz examples confirm the theory. The inferred scheme can tolerate large time-steps and efficiently reproduce the invariant measure.

Many fronts are left open for further investigation. (1) The optimal time-step. We have observed that the inferred schemes perform the best (at reproducing the

invariant measure) when the time-step is medium-large. This observation suggests a trade-off between the 1-step approximation error of the flow map and the accumulated numerical error in the invariant measure. Similar optimality in the medium range was observed in space-time model reduction [26] and in parameter estimation for multiscale diffusion [35]. It is crucial to have a universal *a priori* estimate on the optimal time-step, which can guide general data-driven model reduction approaches. (2) Multi-step noise. We focused on approximate flow maps that use only the increments of the Brownian motion. This limits the performance of the inferred scheme because we omit the details of the stochastic force. A multi-step noise provides the necessary information for further improvements, particularly when the noise is non-stationary [23]. (3) Non-ergodic systems and/or space-time reduction. We expect to extend the framework of ISALT to simulate non-ergodic systems or achieve space-time reduction for high-dimensional nonlinear systems by extracting informed basis functions from the classical numerical schemes.

Acknowledgments. The authors would like to thank the two anonymous reviewers for helpful comments that helped substantially improve the manuscript. FL would like to thank Dr Kevin Lin for helpful discussions.

REFERENCES

[1] Y. Bar-Sinai, S. Hoyer, J. Hickey and M. P. Brenner, Learning data-driven discretizations for partial differential equations, *Proc. Natl. Acad. Sci. USA*, **116** (2019), 15344–15349.
[2] A. J. Chorin and F. Lu, Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics, *Proceedings of the National Academy of Sciences, USA*, **112** (2015), 9804–9809.
[3] A. J. Chorin, F. Lu, R. N. Miller, M. Morzfeld and X. Tu, Sampling, feasibility, and priors in data assimilation, *Discrete Contin. Dyn. Syst.*, **36** (2016), 4227–4246.
[4] W. E, B. Engquist, X. Li, W. Ren and E. Vanden-Eijnden, The heterogeneous multiscale method: A review, In *Commun. Comput. Phys.*, **2** (2007), 367–450.
[5] P. Hall and C. C. Heyde, *Martingale Limit Theory and its Application*, Academic press, 1980.
[6] J. Han, A. Jentzen and W. E, Solving high-dimensional partial differential equations using deep learning, *Proc. Natl. Acad. Sci. USA*, **115** (2018), 8505–8510.
[7] J. A. Hansen and C. Penland, Efficient approximate technique for integrating stochastic differential equations, *Monthly Weather Review*, **134** (2006), 3006–3014.
[8] C. C. Heyde, On the central limit theorem for stationary processes, *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, **30** (1974), 315–320.
[9] Y. Hu, Strong and weak order of time discretization schemes of stochastic differential equatios, In *Séminaire de Probabilités XXX*, Springer, (1996), 218–227.
[10] T. Hudson and X. H. Li, Coarse-graining of overdamped Langevin dynamics via the Mori–Zwanzig formalism, *Multiscale Model. Simul.*, **18** (2020), 1113–1135.
[11] M. Hutzenthaler and A. Jentzen, *Numerical Approximations of Stochastic Differential Equations with Non-globally Lipschitz Continuous Coefficients*, American Mathematical Society, 2015.
[12] M. Hutzenthaler, A. Jentzen and P. E. Kloeden, Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients, *Ann. Appl. Probab.*, **22** (2012), 1611–1641.
[13] A. Jentzen and P. Kloeden, Taylor expansions of solutions of stochastic partial differential equations with additive noise, *Ann. Probab.*, **38** (2010), 532–569.
[14] S. W. Jiang and J. Harlim, Modeling of missing dynamical systems: Deriving parametric models using a nonparametric framework, *Res. Math. Sci.*, **7** (2020), Paper No. 16, 25 pp.
[15] R. Khasminskii, *Stochastic Stability of Differential Equations*, volume **66**. Springer-Verlag Berlin Heidelberg, 2nd edition, 2012.
[16] B. Khouider, A. J. Majda and M. A. Katsoulakis, Coarse-grained stochastic models for tropical convection and climate, *Proc. Natl. Acad. Sci. USA*, **100** (2003), 11941–11946.

[17] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 3rd edition, 1992.
[18] K. Law, A. Stuart and K. Zygalakis, *Data Assimilation: A Mathematical Introduction*, Springer, 2015.
[19] F. Legoll and T. Lelièvre, Effective dynamics using conditional expectations, *Nonlinearity*, **23** (2010), 2131–2163.
[20] F. Legoll, T. Leliévre and U. Sharma, Effective dynamics for non-reversible stochastic differential equations: A quantitative study, *Nonlinearity*, **32** (2019), 4779–4816.
[21] H. Lei, N. A. Baker and X. Li, Data-driven parameterization of the generalized Langevin equation, *Proc. Natl. Acad. Sci. USA*, **113** (2016), 14183–14188.
[22] B. Leimkuhler and C. Matthews, *Molecular Dynamics*, Springer, 2015.
[23] Y. Li and J. Duan, A data-driven approach for discovering stochastic dynamical systems with non-Gaussian Lévy noise, *Phys. D*, **417** (2021), 132830, 12 pp.
[24] K. K. Lin and F. Lu, Data-driven model reduction, Wiener projections, and the Koopman-Mori-Zwanzig formalism, *J. Comput. Phys.*, **424** (2021), 109864, 33 pp.
[25] S. Liu, L. Grzelak and C. W. Oosterlee, The seven-league scheme: Deep learning for large time step monte carlo simulations of stochastic differential equations, arXiv:2009.03202, (2020).
[26] F. Lu, Data-driven model reduction for stochastic Burgers equations, *Entropy*, **22** (2020), Paper No. 1360, 22 pp.
[27] F. Lu, K. K. Lin and A. J. Chorin, Comparison of continuous and discrete-time data-based modeling for hypoelliptic systems, *Commun. Appl. Math. Comput. Sci.*, **11** (2016), 187–216.
[28] F. Lu, K. K. Lin and A. J. Chorin, Data-based stochastic model reduction for the Kuramoto–Sivashinsky equation, *Phys. D*, **340** (2017), 46–57.
[29] F. Lu, M. Maggioni and S. Tang, Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories, *J. Mach. Learn. Res.*, **22** (2021), Paper No. 32, 67 pp.
[30] F. Lu, M. Zhong, S. Tang and M. Maggioni, Nonparametric inference of interaction laws in systems of agents from trajectory data, *Proc. Natl. Acad. Sci. USA*, **116** (2019), 14424–14433.
[31] Y. Maday and G. Turinici, A parareal in time procedure for the control of partial differential equations, *C. R. Math. Acad. Sci. Paris*, **335** (2002), 387–392.
[32] A. J. Majda and J. Harlim, Physics constrained nonlinear regression models for time series, *Nonlinearity*, **26** (2013), 201–217.
[33] X. Mao, *Stochastic Differential Equations and Applications*, Elsevier, 2007.
[34] J. C. Mattingly, A. M. Stuart, and D. J. Higham, Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise, *Stochastic Process. Appl.*, **101** (2002), 185–232.
[35] G. A. Pavliotis and A. M. Stuart, Parameter estimation for multiscale diffusions, *J. Statist. Phys.*, **127** (2007), 741–781.
[36] G. O. Roberts and R. L. Tweedie, Exponential convergence of Langevin distributions and their discrete approximations, *Bernoulli*, **2** (1996), 341–363.
[37] W. Rümelin, Numerical treatment of stochastic differential equations, *SIAM J. Numer. Anal.*, **19** (1982), 604–613.
[38] J. Sirignano and K. Spiliopoulos, DGM: A deep learning algorithm for solving partial differential equations, *J. Comput. Phys.*, **375** (2018), 1339–1364.
[39] L. Yang, D. Zhang and G. E. Karniadakis, Physics-informed generative adversarial networks for stochastic differential equations, *SIAM J. Sci. Comput.*, **42** (2020), A292–A317.

Received February 2021; revised June 2021; early access September 2021.

E-mail address: xli47@uncc.edu
E-mail address: feilu@math.jhu.edu
E-mail address: xye2@albany.edu