Das. *Complex Eng. Syst.* **2025**, *5*, 15
DOI: 10.20517/ces.2025.59

Complex Engineering Systems

Research Article

Open Access

A probabilistic approach to drift estimation from stochastic data

Suddhasattwa Das

Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA.

Correspondence to: Dr. Suddhasattwa Das, Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409, USA. E-mail: suddas@ttu.edu

How to cite this article: Das, S. A probabilistic approach to drift estimation from stochastic data. *Complex Eng Syst* **2025**, *5*, 15. http://dx.doi.org/10.20517/ces.2025.59

Received: 26 Aug 2025 **First Decision:** 28 Sep 2025 **Revised:** 29 Sep 2025 **Accepted:** 15 Oct 2025 **Published:** 7 Nov 2025

Academic Editor: Duxin Chen **Copy Editor:** Fangling Lan **Production Editor:** Fangling Lan

Abstract

Time series generated from a dynamical source can often be modeled as sample paths of a stochastic differential equation. The time series thus reflects the motion of a particle that flows along the direction provided by a drift/vector field, and is simultaneously scattered by the effect of white noise. The resulting motion can only be described as a random process instead of a solution curve. Due to the non-deterministic nature of this motion, the task of determining the drift from data is quite challenging, since the data does not directly represent the directional information of the flow. This paper describes an interpretation of a drift as a conditional expectation, which makes its estimation feasible via kernel-integral methods. In addition, some techniques are proposed to overcome the challenge of dimensionality if the stochastic differential equations carry some structure enabling sparsity. The technique is shown to be convergent, consistent and permits a wide choice of kernels.

Keywords: Markov kernel, drift estimation, compact operators, reproducing kernel Hilbert space

1. INTRODUCTION

The dynamics present in several physical phenomena are governed by both deterministic laws of motion and stochastic driving terms. Stochastic differential equations (SDEs) provide a common mathematical description

for these phenomena. A general d-dimensional SDE takes the form

$$dX(t) = V\left(X(t)\right) + G\left(X(t)\right) dW^t, \tag{1}$$

in which the term V is known as the *drift*, which is a function of the current state. The component V plays the role of a vector field over the state space $\mathbb{R}^d$, providing a deterministic direction to the flow at all locations. The term G is known as *diffusion*. It relays the effect of an external white noise onto the motion. Such SDEs have been used effectively in the study of a wide range of phenomena, such as control systems[1], electrical systems[2,3], general mass transport[4], and fluid dynamics[5,6].

This article focuses on the data-driven discovery of SDEs. A convergent and consistent technique is presented for estimating the drift term V in SDEs, from data collected from observations of the system at regular sampling intervals of Δt. The drift is a vector field and thus a deterministic function of the state. However, due to the stochasticity originating from white noise, the state at time t of an SDE-driven system is not a deterministic function of its current state. Rather, it has a probability distribution conditioned on the current state. As a result, the solution $X(t, x_0)$ when expressed as a function of time and initial state x_0 is not a deterministic curve but a Markov process in which almost every sample path is continuous but non-differentiable[7,8]. This stochasticity and lack of differentiability makes the task of drift estimation challenging, as the drift is essentially a differential operator.

The Weiner process that drives an SDE of the form (1) can also be interpreted as a probability space, whose domain is the *path-space* $\mathbf{\Omega}_{\text{path}}$ of all possible random Brownian walks. The space $\mathbf{\Omega}_{\text{path}}$ provides a concrete way to identify the source of stochasticity in the solutions of the SDE. Every random point from $\mathbf{\Omega}_{\text{path}}$ corresponds to a particular sample of a Brownian path, and thus corresponds to a unique realization of (1). This path space formalism of stochastic processes, championed by Doob[9] and Arnold[7], clarifies the source of stochasticity in the study of SDEs. The path space $\mathbf{\Omega}_{\text{path}}$ itself is usually left undetermined, and the SDE is studied via the distributions it induces on $\mathbb{R}^d$.

Drift estimation of SDE time series requires a treatment completely different from those used for numerical differentiation. Figure 1 provides an outline of the mathematical problem, and the proposed method of converting it into a task of conditional expectation. In the usual probabilistic approach to SDEs, the initial state x_0 to the SDE (1) is not fixed but is itself a random variable X_0. This stochasticity in the location of x_0 is propagated into the future states in combination with the effect of white-noise. As a result, the state at any time t is a random variable X_t. This 1-parameter family $\{X_t \ : \ t \geq 0\}$ of random variables is a Markov process. Its key feature is the useful *memoryless property* - for any three time instants $t_1 < t_2 < t_3$, the distribution of X_{t_3} can be determined from the value of X_{t_2} alone, and may be treated as independent of X_{t_1}. Figure 2 presents some examples of SDEs simulated using the Euler-Maruyama method. We make the following assumption:

Assumption 1 *The SDE* (1) *has a stationary probability measure* μ_{stat}.

Stationarity of μ_{stat} means that if X_0 is distributed according to μ_{stat}, then so will be X_t for every $t \geq 0$. Stationary measures are the stochastic analog of *ergodic measures* in deterministic dynamical systems, and are one of the pillars of data-driven studies of stochastic systems. Stationarity is empirically observed in most systems, although it has been theoretically proven only under certain sets of conditions[10–12]. Assumption 1 guarantees that the solution $X(t)$ is a stationary Markov process. This means that the conditional distribution of X_{t_2} with respect to (w.r.t.) its location at a time $t_1 < t_2$, only depends on the time elapsed $t_2 - t_1$.

A Markov process can be alternatively studied by the continuous transformation it induces on probability measures. Recall that the distribution of X_t is determined by that of X_0. If the distribution of X_0 is absolutely continuous w.r.t. the Lebesgue volume measure on $\mathbb{R}^d$, then the same will be true for X_t for every $t > 0$. Thus, an SDE induces a dynamics on the space of probability densities as well. This dynamics or continuous

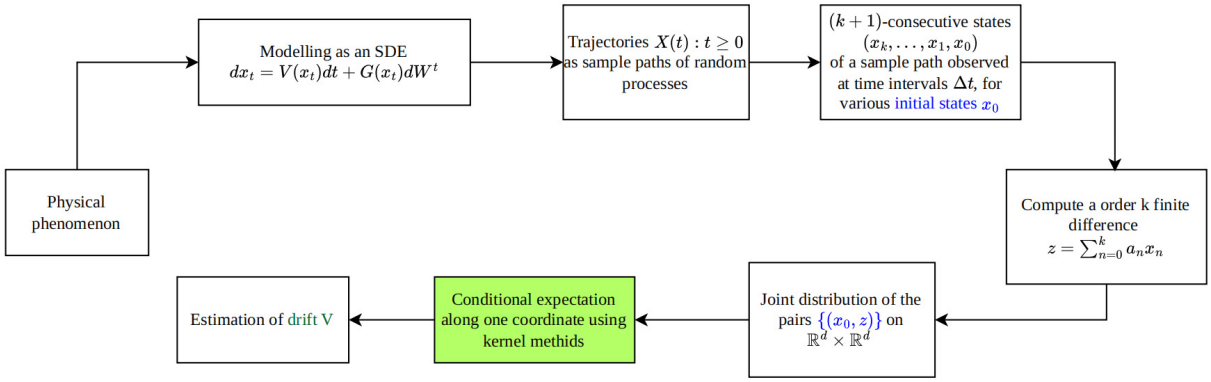

Figure 1. Outline of the theory. Many dynamical processes in nature can be modeled as an SDE, which has a drift and diffusion component. Unlike deterministic systems, the consecutive points on a trajectory are not related deterministically, but have a distribution. The flowchart shows how we can collect $(k+1)$-length snapshots of this process, and apply purely data-driven algorithms to estimate the drift. No presumption is needed on a prior distribution on the coordinates, or on a parametric form for the SDE terms. The article presents the details of these methods, their theoretical foundations in Probability theory, and a rigorous demonstration that they provide a convergent, unbiased estimate.

transformation of the densities can be tracked by the *forward Kolmogorov operator* $\mathcal{L}$. This second order differential operator whose action on a C^2 (Recall that a function is called C^r if it is r-times differentiable, and its first r derivatives are continuous). function g is given by[7]

$$(\mathcal{L}g)(X,t) := \sum_{i=1}^{d} V(x)_i \frac{\partial}{\partial_i} g(x) + \sum_{i,j=1}^{d} \left(\frac{\partial^2}{2\partial_i \partial_j} g \right)(x) A_{i,j}(x), \tag{2}$$

where $A(x) = 0.5G(x)G(x)^T$ is the $d \times d$ covariance matrix. One of the pillars of the theory of Markov processes is that the derivative of the conditional expectations of a stationary Markov process is provided by the generator $\mathcal{L}$ (see[13] Cor 160):

$$\frac{d}{ds}|_{s=t} \mathbb{E}\left(\phi\left(X(s)\right) \mid X(t)\right) = \mathcal{L}X(t), \qquad \forall \phi \in \mathrm{dom}(\mathcal{L}). \tag{3}$$

Equation (3) is a combination of a second order differential operator, the measure theoretic operation of conditional expectation, and a time derivative. It provides a bridge between statistical properties (i.e., the distribution of the data), and differential properties (i.e., the drift). Of special interest to us is the particular instance of (3) when $\phi = \mathrm{Id}_{\mathbb{R}^d}$ and $t = 0$. In that case, we get

$$\lim_{t \to 0^+} \frac{1}{t} \mathbb{E}\left(X(t, x_0) - x_0\right) = \left(\mathcal{L}\,\mathrm{Id}_{\mathbb{R}^d}\right)(x_0) = V(x_0). \tag{4}$$

Equation (4) is the core mathematical principle of our numerical technique. In fact, we use a third order version of (4) given by

$$\Delta t V(x_0) = 18\mathbb{E}\left(X\left(\Delta t, x_0\right) - x_0\right) - 9\mathbb{E}\left(X\left(2\Delta t, x_0\right) - x_0\right) + 2\mathbb{E}\left(X\left(3\Delta t, x_0\right) - x_0\right) + O\left(\Delta t^3\right). \tag{5}$$

This identity is based on a Taylor series expansion of operator $\mathcal{L}$:

$$\frac{1}{\Delta t}\left[\phi(X_{t+\Delta t}) - \phi(X_t)\right] = \sum_n \frac{1}{n!} \mathcal{L}^n \phi(X_t) \left(\Delta t\right)^n.$$

The precise nature of the convergence depends upon the regularity of V and G. See Eq. 21 in ref.[14] for more details. The conditional expectations here are taken w.r.t the path-space Ω_{path}. Equations (4) and (5) provide a connection between successive states of an SDE path, and the drift term of the SDE. Thus, the task of estimating

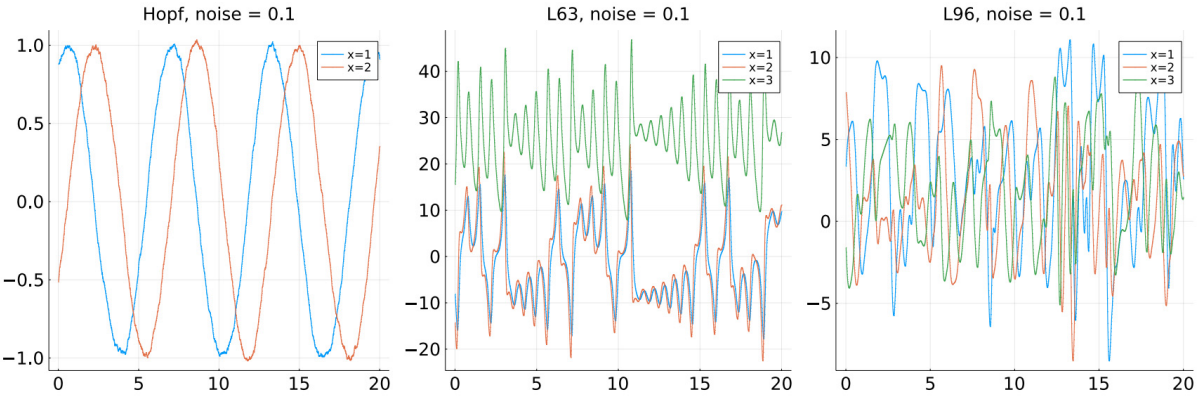

Figure 2. Orbits from SDEs. The three plots show sample paths of SDEs, in which the drifts are respectively the Hopf oscillator (27), Lorenz-63 system (26) and the Lorenz 96 system (28). The SDEs have a diagonal diffusion term given by (29). The parameters of the experiments are provided in Table 1. The legends indicate the index of the coordinates being plotted. The paper presents a numerical method for extracting the drift from such time series generated by SDEs. The results of the reconstruction have been displayed in Figures 3, 4, 5, respectively.

the drift becomes the task of estimating certain conditional expectations. This means that the determination of a deterministic, pre-defined vector valued function can be placed in a probabilistic context, and solved by probabilistic means.

A careful design of a procedure for this task requires a careful understanding of the nature of the data. Formally, we assume the following about the data:

Assumption 2 *There is sampling interval* $\Delta t > 0$*, a collection of states* $x_1, \ldots, x_N$ *equidistributed w.r.t. the stationary measure* μ_{stat}*, and a collection of* N *4-tuples of points*

$$\{(X(3\Delta t, x_n), X(2\Delta t, x_n), X(\Delta t, x_n), x_n) \ : \ 1 \leq n \leq N\} \tag{6}$$

Thus, our assumption on the data is that we have snapshots of sample trajectories of length at least 4, and that the data is equidistributed w.r.t. the unknown stationary measure. The assumption of *equidistribution* is essential and fundamental to most data-driven methods for discovering dynamical systems. Our assumption on data requires neither a single orbit, nor long sequences of data. It does allow the points x_n to be gathered from a single trajectory of the SDE, observed at various time instants. This makes our technique applicable to systems that are being intermittently observed.

Our goal can now be precisely stated to be the estimation of the drift V from (1), while working under the constraints and conditions of Assumptions 1 and 2. Although there have been many a wide range of techniques for approximating the drift for ordinary differential equations (ODEs)[15–19], it is still a challenge to develop a technique that fulfills the following criteria:

(i) The technique is nonparametric; i.e., it does not assume a prescribed parametric format, such as Gaussian processes[20] or semi-parametric forms such as linear diffusion terms[21]
(ii) The technique is convergent with data.
(iii) The technique does not assume a prior distribution; such as Gaussian priors[22,23].
(iv) The technique is adaptable to high dimensional systems.

Effective nonparametric techniques have been built[24,25] which are well suited to sparse sampling but rely on the assumption of prior distributions. Another important work[26] considers data collected from SDE sampled at uneven intervals. Besides the probabilistic approach presented in this article, four other notable and distinct techniques deserve special mention - a model reduction technique of Ye *et al.*[27]; the one based

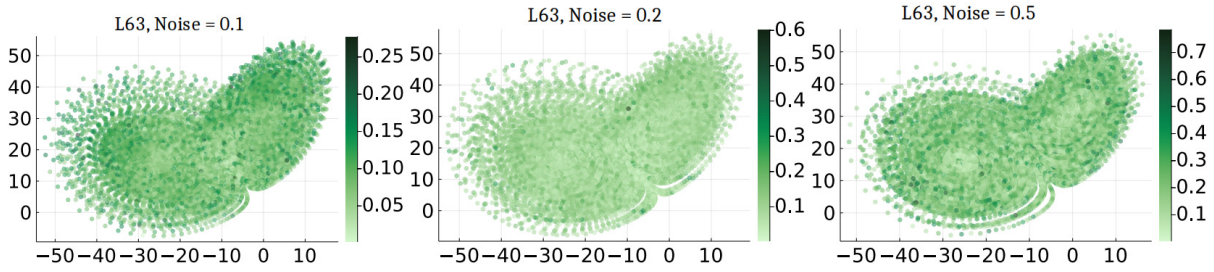

Figure 3. Estimating the drift in a stochastic Lorenz 63 model. The three figures present a heatmap of the pointwise error in computing the drift (26). The colored area of the phase space represents the support of the stationary measure of the underlying SDE.

on the *Magnus expansion*[28]; a back-stepping technique that converts the nonlinear process into an Ornstein-Uhlenbeck (OU) process[29]; and the *multi-level parameter estimation network* for parametric SDEs with jump noises. Our technique achieves all of the first three objectives and overcomes the challenge of dimensionality in vector fields structured in a certain way, as explained later in Section 5.

Our method is based on a measure theoretic insight. Set two spaces

$$\mathcal{X} = \mathbb{R}^d, \quad \mathcal{Y} = \mathbb{R}^{3d}.$$

These two spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively represent the spaces from which x_0 and the triples $(X(\Delta t, x_0), X(2\Delta t, x_0), X(3\Delta t, x_0))$ are drawn. Define a function

$$f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d, \quad x_0 \times (x_1, x_2, x_3) \mapsto \frac{1}{\Delta t}\left[18x_1 - 9x_2 + 2x_3 - 11x_0\right] \tag{7}$$

Due to the stationarity assumption from Assumption 1, there is a natural measure μ on $\mathcal{X} \times \mathcal{Y}$ representing the joint distribution

$$(x_0, X(\Delta t, x_0), X(2\Delta t, x_0), X(3\Delta t, x_0)) \sim \mu$$

Let $\text{proj}_1 : \mathcal{X} \times \mathcal{Y} \to \mathcal{X}$ denote the projection map to the first component. Then (5) can be interpreted as stating that:

$$V = \mathbb{E}_\mu\left(f \mid \text{proj}_1\right) + O\left(\Delta t^3\right) \tag{8}$$

Thus, the task of estimating the drift has been transformed into the task of finding the conditional estimation of one projection function w.r.t. another. The conditional expectation is reconstructed or learnt using an operator theoretic tool from[30]. It will be a combination of kernel-based techniques and the third order approximation of (8). This completes a description of the problems we aim to solve, and the underlying mathematical principles.

Outline We next present in Section a numerical recipe for implementing the conditional expectation route to drift estimation. Subsequently, in Section , we discuss the convergence of our methods, and identify the dependence of the convergence on factors such as data size and noise levels. Then, in Section , we apply the technique to some common examples from Dynamical systems theory. The examples reveal the challenges posed by high dimensional systems. We propose some adjustments to the algorithms to overcome this practical challenge in Section . We end with some discussions in Section , summarizing our results and discussing new challenges.

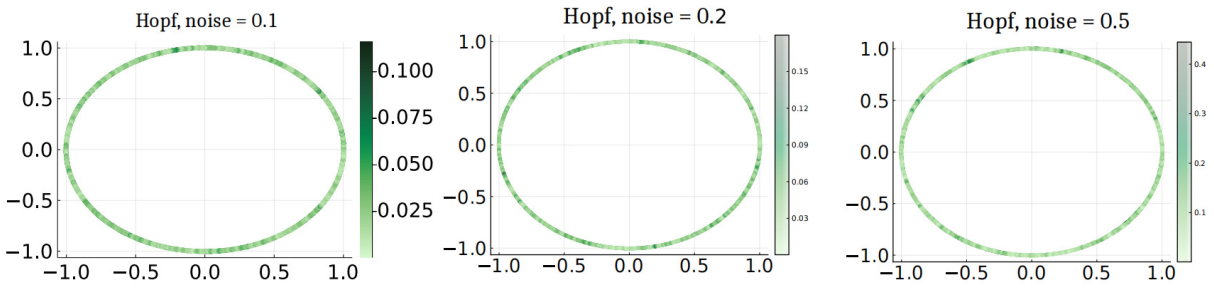

Figure 4. Estimating the drift in a stochastic version of the Hopf oscillator (27). The images carry the same interpretation as Figure 3.

2. THE TECHNIQUE

Our technique makes multiple and different uses of kernels and kernel integral methods. A *kernel* on a space $\mathcal{Z}$ is a bivariate function $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$, and is interpreted to be a measure of similarity between points on $\mathcal{Z}$. Kernel-based methods offer a nonparametric approach to learning, and have been used with success in many diverse fields such as spectral analysis [31,32], discovery of periodic and chaotic components of various real world systems [33,34], and even abstract operator valued measures [35]. The kernels we used are based on *Gaussian* kernels, defined as

$$k_{\text{Gauss},\epsilon}(x, y) := \exp\left(-\frac{1}{\epsilon}\,\text{dist}(x, y)^2\right), \quad \forall x, y \in \mathcal{Z}. \tag{9}$$

The bandwidth ϵ controls how quickly the value of the kernel decays to zero as x and x' grow apart. We use a Markov normalized version of the Gaussian kernel (9). Given a measure β on $\mathcal{Z}$ and a bandwidth parameter $\epsilon > 0$, we can create a new kernel

$$k^{\beta}_{\text{Gauss},\epsilon} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}, \quad k^{\beta}_{\text{Gauss},\epsilon}(x, x') := k_{\text{Gauss},\epsilon}(x, x') \Big/ \int_{\mathcal{Z}} k_{\text{Gauss},\epsilon}(x, x'')d\beta(x''). \tag{10}$$

The kernel $k^{\beta}_{\text{Gauss},\epsilon}$ has the property that for every x, the integral w.r.t. the second variable x' is 1. Thus, the kernel $k^{\beta}_{\text{Gauss},\epsilon}$ mimics a discrete time Markov process whose state space is $\mathcal{Z}$. Closely associated with kernels are kernel integral operators. Now suppose that $\mathcal{X}$ is a C^r-smooth manifold and k is a C^r-smooth kernel. Given a probability measure α on $\mathcal{Z}$, one has the following integral operator associated to k:

$$K^{\alpha} : L^2(\alpha) \to C^r(\mathcal{X}), \quad (K^{\alpha}\phi)(x) := \int_{\mathcal{X}} k(x, y)\phi(y)d\nu(y).$$

Thus, if the kernel k is C^r, it converts generally non-smooth $L^2(\alpha)$ functions into C^r functions. For this reason, kernel integral operators are also known as smoothing operators. We also require our kernel to be *strictly positive definite*, which means that given any distinct points $x_1, \ldots, x_N$ in $\mathcal{Z}$, numbers $a_1, \ldots, a_N$ in $\mathbb{R}$, the sum $\sum_{i=1}^{N}\sum_{i=1}^{N} a_i a_j k(x_i, x_j)$ is non-negative, and zero iff all the a_i-s are zero.

We next present a useful notation: Let $\theta > 0$ be a fixed constant. Given a generic linear regression problem $Ax = y$, the θ-regularized least squares solution will be the vector

$$x_{ls,\theta-reg} := \left[A^T A + \theta\right]^{-1} A^T y. \tag{11}$$

We now present our first algorithm from [30] for computing conditional expectation from data.

Algorithm 1 *Kernel base estimation of conditional expectation.*

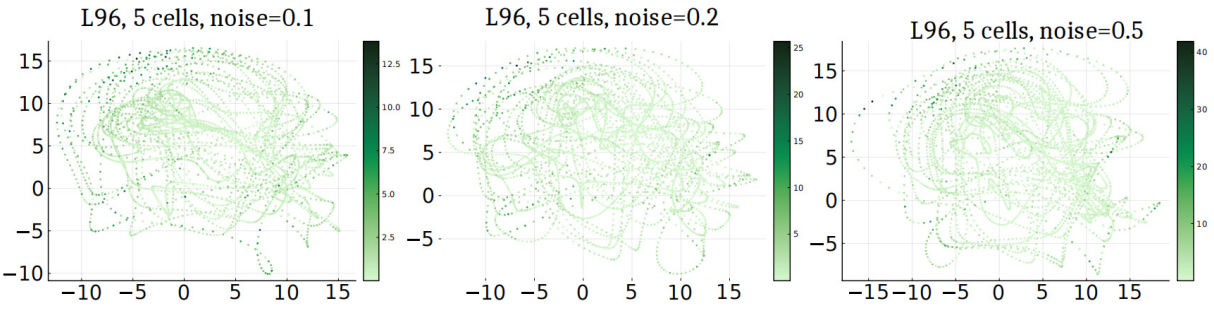

Figure 5. Estimating the drift in a stochastic version of the L96 system (28) with 5 oscillators. The images carry the same interpretation as Figure 3.

- ***Input.*** *A sequence of pairs* $\{(x_n, y_n) \ : \ n = 1, \dots, N\}$ *with* $x_n \in \mathbb{R}^d$ *and* $y_n \in \mathbb{R}$.
- ***Parameters.***
  1. *Strictly positive definite kernel* $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$.
  2. *Smoothing parameter* $\epsilon_1 > 0$.
  3. *Sub-sampling parameter* $M \in \mathbb{N}$ *with* $M < N$.
  4. *Ridge regression parameter* θ.
- ***Output.*** *A vector* $\vec{a} = (a_1, \dots, a_M) \in \mathbb{R}^M$ *that creates a function*

$$\phi(x) \approx \sum_{m=1}^{N} a_m k(x, x_m), \qquad \forall x \in \mathbb{R}^d,$$

  which approximates the conditional expectation of the y-variable w.r.t. the x variable.
- ***Steps.***
  1. *Compute a Gaussian Markov kernel matrix using* (10)*:*

$$\left[G_{\epsilon_1}\right] \in \mathbb{R}^{N \times N}, \qquad \left[G_{\epsilon_1}\right]_{i,j} = k^{\beta}_{Gauss,\epsilon}(x_i, x_j). \tag{12}$$

  2. *Compute a Markov kernel* $[P] \in \mathbb{R}^{N \times N}$ *as* $[P]_{i,j} := p(x_i, x_j)$
  3. *Compute the kernel matrix* $[K] \in \mathbb{R}^{N \times M}$ *as* $[K]_{i,j} = k(x_i, x_j)$.
  4. *Find a vector* $\vec{a} \in \mathbb{R}^M$ *as the* θ*-regularized least-squares solution to the equation*

$$[P]\,[K]\,\vec{a} = [P]\left[G_{\epsilon_1}\right]\vec{y}. \tag{13}$$

Algorithm 1 has two components, the choice of a reconstruction kernel k, and a Markov kernel p which approximates the smoothing operator. We usually choose p to be the Markov normalized Gaussian kernel from (10). The algorithm is convergent in the following sense:

Lemma 0.1 *([30] Thm 2) Suppose there is a probability space* (Ω, μ) *and two measurable functions* $X : \Omega \to \mathbb{R}^d$ *and* $Y : \Omega \to \mathbb{R}$. *Suppose that the sequence* $\{(x_n, y_n)\}_{n \in \mathbb{N}}$ *is equidistributed w.r.t. some measure on* $\mathbb{R}^d \times \mathbb{R}$. *Let the conditional expectation* $V = \mathbb{E}_\mu(Y|X)$ *lie within the reproducing kernel Hilbert space (RKHS) spanned by* k. *Then, the output* $\vec{a}$ *of Algorithm 1 satisfies:*

$$\lim_{M \to \infty} \lim_{N \to \infty, \theta \to 0^+} \left\| \sum_{n=1}^{N} a_n k(\cdot, x_n) - \mathcal{G}^{\mu}_{\epsilon_1} V \right\|_{\mathcal{H}} = 0, \qquad \lim_{\epsilon_1 \to 0+} \left\| \mathcal{G}^{\mu}_{\epsilon_1} V - V \right\|_{L^2(\nu)} = 0. \tag{14}$$

Algorithm 1 is a general algorithm for computing conditional expectation, and is at the core of our numerical procedure. We make an important note here that the application of the Gaussian smoothing $\left[G_{\epsilon_1}\right]$ is optional, and does not affect the analytic properties of the convergence scheme. We next show its use in drift estimation. Given any $d \times M$ matrix A, we use A_i and A^j to denote its i-th column vector and j-th row vector, respectively.

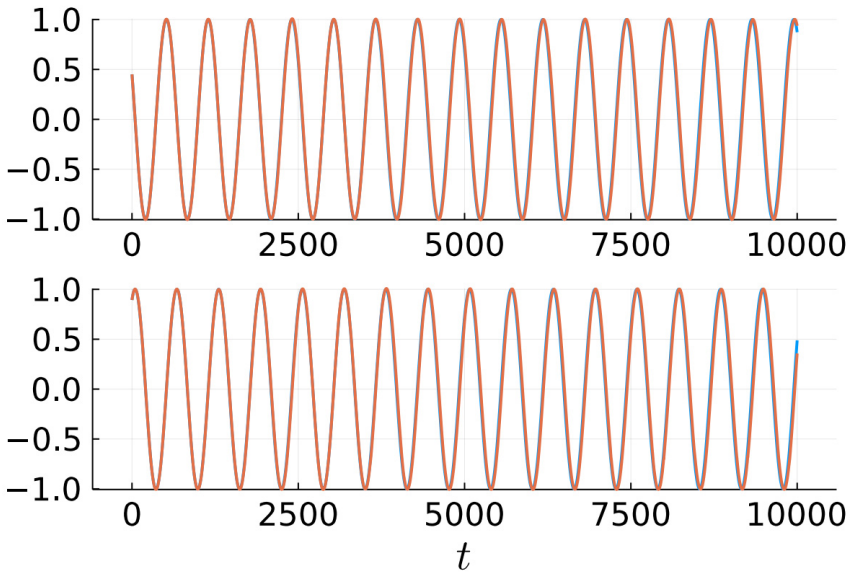

Figure 6. Reconstruction of the Hopf oscillator. The trajectories corresponding to the true and reconstructed drifts are in blue and orange, respectively. Due to a strong match in the reconstructed drifts, and the stability of rotational dynamics, the orbits overlap.

Given a sequence $\{z_n\}_{n=1}^N$ of d-dimensional vectors, we use $\left\{z_n^{(i)}\right\}_{n=1}^N$ to denote the 1-dimensional time series created by extracting the i-th coordinate from these vectors. We now present how Algorithm 1 can be used to estimate the drift from an SDE orbit.

Algorithm 2 *Drift estimation as a conditional expectation*

- ***Input.*** *Dataset* $\{(X(3\Delta t, x_n), X(2\Delta t, x_n), X(\Delta t, x_n), x_n)\}_{n=1}^N$ *as in Assumption 2.*
- ***Parameters.*** *Same as Algorithm 1.*
- ***Output.*** *A matrix* $A \in \mathbb{R}^{d\times M}$ *representing a drift*

$$\hat{V}(X) \approx \sum_{m=1}^{M} A_m k(x, x_m), \qquad \forall x \in \mathbb{R}^d.$$

- ***Steps.***
  1. *Compute*

$$z_n = \frac{1}{\Delta t}\left[18X(\Delta t, x_n) - 9X(2\Delta t, x_n) + 2X(3\Delta t, x_n) - 11x_n\right], \qquad \forall n \in 1, \ldots, N.$$

  2. *For each* $i \in 1, \ldots, d$, *pass the time series* $\left\{\left(x_n, z_n^{(i)}\right)\right\}_{n=1}^N$ *into Algorithm 1, along with the rest of the parameters.*
  3. *Store the result in a m-dimensional column vector* $\vec{a}^{(i)}$.
  4. *Compute the* $d \times M$ *matrix A such that*

$$A^T = \left[\vec{a}^{(1)}, \vec{a}^{(2)}, \ldots, \vec{a}^{(d)}\right].$$

The convergence of Algorithm 2 is guaranteed by Lemma 0.1 and (8). We investigate this convergence more closely in the next section.

3. CONVERGENCE ANALYSIS

In this section, we conduct a precise analysis on the convergence of our results and its dependence on various algorithmic and data-source parameters. Throughout this section we assume that Assumptions 1 and 2 hold;

the drift and diffusion terms in (1) are continuous functions; and that $a_{N,M,\theta}$ is the output of Algorithm 2 with N data points, sub-sampling parameter M, and regression regularization parameter θ. We use $\hat{V}_{N,M,\theta}$ to denote the resulting approximation of the drift.

Theorem 1 (Almost sure convergence) *Suppose Assumptions 1 and 2 hold. If the drift V lies in the RKHS $\mathcal{H}$, then the following convergence holds almost surely:*

$$\boxed{\lim_{M\to\infty}\lim_{N\to\infty,\theta\to 0^+}\left\|\hat{V}_{N,M,\theta}-V\right\|_{\sup}\overset{a.s.}{=}0.} \tag{15}$$

Theorem 1 will be shown to be a consequence of functional analytic results from the work by Das[30]. The convergence result does not reveal the dependence of the rate of convergence on the choice of parameters. It also relies on the assumption that V lies in the hypothesis space $\mathcal{H}$. The three main parameters involved are:

(i) the number of data-samples N, which dictates how well the distribution of the data points in $\mathbb{R}^{4d}$ is approximated;
(ii) choice of the sub-sampling parameter M;
(iii) a regularization parameter $\theta > 0$, as explained in (11), for the linear regression problem.

The next theorem presents a more nuanced view of the convergence, in terms of error bounds:

Theorem 2 (Parameter selection within error bounds) *Suppose Assumptions 1 and 2 hold. Fix error bounds $\delta_1, \delta_2, \delta_3 > 0$ and an uncertainty bound $\delta_4 > 0$.*

(i) For M large enough, V may be approximated uniformly within an error of δ_1 as a sum of kernel sections

$$\left\|V-\sum_{m=1}^{M}\bar{a}_m k\left(\tilde{x}_m,\cdot\right)\right\|_{\sup\mathcal{X}}<\delta_1.$$

Here, $\{\tilde{x}\}_{m=1}^{M}$ is a collection of M points from the dataset assumed in Assumption 2.

(ii) Given such a choice of M, there exists $\theta_0 > 0$ and $N_0' \in \mathbb{N}$ such that for every $N > N_0'$ and $\theta \in (0, \theta_0)$, if the diffusion term $G \equiv 0$, then the approximation error of the approximate drift $\sum_{m=1}^{M}\bar{a}_m k\left(\tilde{x}_m,\cdot\right)$ is less than δ_2.

(iii) Fix such a $\theta \in (0, \theta_0)$. Then there is an N_0 satisfying

$$N_0 > N_0', \qquad N_0 = O\left(\frac{d^2}{\delta_3^4\delta_4^2}\right),$$

such that for every $N > N_0$, with probability at least $1-\delta_4$, the total error in approximation is uniformly bounded by $\delta_1+\delta_2+\delta_3$.

Theorem 2 also presents the order in which the parameters are to be tuned, starting with M and ending in N. The error term δ_1 is related to the approximation error in functional space. The error term δ_2 is controlled by the regression parameter θ. The error term δ_3 is a result of the fluctuations of cumulative error terms from the mean value, as will be described later. The error term δ_4 is just a bound on uncertainty. The gross error $\delta_1+\delta_2+\delta_3$ thus reflects an accumulation of errors from three separate levels of approximation. Theorem 2 delineates the contribution of each parameter choice to the error. Note the appearance of d^2 in the last lower bound. This term reveals the inevitable challenge of dimensionality.

The proofs of both theorems emerge from a careful comparison of the output of the algorithm against the true drift. Recall the spaces $\mathcal{X}, \mathcal{Y}$, measure μ and functions proj_1, f from Section . Note that the function f is just the transform on the datapoints, used in Algorithm 2. Let β denote the empirical measure borne by the N data points used in algorithm 2, and β_X denote the projection of β to $\mathcal{X}$. Let ν denote the empirical measure

created from a sub-sampling of β_X. These notations allow us to interpret the various objects in the numerical procedure in a rigorous operator theoretic setting. With this in mind the regression (13) becomes

$$V \approx \hat{V}_{\beta,\nu,\theta}, \quad \hat{V}_{\beta,\nu,\theta} := K^\nu a_{\beta,\nu,\theta}, \quad a_{\beta,\nu,\theta} := \underset{\theta-\text{reg}}{\arg\min} \left\{ \left\| \mathbb{P}^{\beta_X} K^\nu a - \tilde{\mathbb{P}}^\beta f \right\|_{L^2(\nu)} \; : \; a \in L^2(\nu) \right\}. \tag{16}$$

Algorithm 1 implements the least-squares regression problem posed in (16). The measures β and β_X are completely characterized by the data count N, ν is characterized by a sub-sampling parameter M. Recall that given any measure α on $\mathcal{X}$, we use P^α and K^α to denote the resulting kernel integral operators with kernels p and k respectively. Moreover, we shall use $\tilde{p}$ to denote the trivially extended kernel function

$$\tilde{p} : \mathcal{X} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}, \quad (x_1, x_2, y) \mapsto p(x_1, x_2).$$

Similarly, given any measure $\tilde{\alpha}$ on $\mathcal{X} \times \mathcal{Y}$, the resulting integral operator with kernel $\tilde{p}$ will be denoted as $\tilde{P}^\alpha$. This notation allows the output $\hat{V}_{\beta,\nu,\theta}$ of Algorithm 1 to be expressed concisely as

$$\begin{aligned} \hat{V}_{N,M,\theta} := \hat{V}_{\beta,\nu,\theta} &:= [K^\nu] \left[[K^\nu]^T \left[P^{\beta_X}_{\epsilon_3} \right]^T \left[P^{\beta_X}_{\epsilon_3} \right] [K^\nu] + \theta \right]^{-1} [K^\nu]^T \left[P^{\beta_X}_{\epsilon_3} \right]^T \left[\tilde{P}^{\beta}_{\epsilon_3} \right] f \\ &= [K^\nu] \left[A_{N,\epsilon_3,M} + \theta \right]^{-1} [K^\nu] \left[B_{N,\epsilon_3} \right] f \end{aligned} \tag{17}$$

where we have set

$$B_{N,\epsilon_3} := \left[P^{\beta_X}_{\epsilon_3} \right]^T \left[P^{\beta_X}_{\epsilon_3} \right], \quad A_{N,\epsilon_3,M} := [K^\nu] \, B_{N,\epsilon_3} \, [K^\nu] \, .$$

Each of the matrices represented in square brackets above is also a kernel integral operator w.r.t. the finite empirical measures β, ν and β_X. Equation (17) reveals that the overall transformation is a sequence of linear transformations. Lemma 0.1 and (8) are direct statements about the almost sure convergence of the $\hat{V}_{\beta,\nu,\theta} = \hat{V}_{N,M,\theta}$ above, thus ensuring the almost-sure convergence (15) expressed in Theorem 1. We now move on to Theorem 2.

Separating noise Since we have replaced μ with β and μ_X with ν, the equality (8) is no longer satisfied. In that case, the following difference becomes relevant:

$$\eta_\beta := \left(\tilde{P}^\beta - \tilde{P}^\mu \right) f + \left(P^{\beta_X} - P^{\mu_X} \right) V. \tag{18}$$

In the limit of infinite data, $\beta = \mu$ and $\beta_X = \mu_X$, so the difference η_β becomes zero. We can now write

$$\tilde{P}^\beta f = P^{\beta_X} V + \eta_\beta. \tag{19}$$

The left-hand side (LHS) of (19) is the right-hand side (RHS) of the regression problem being solved in (16) and (17). Thus, (19) expresses the RHS of the regression problem as an integral transform $P^{\beta_X} V$ of the true field V, along with the noise effect η_β defined in (18). Now note the exact expression for η_β:

$$\left(\eta_\beta \right)_j = \frac{1}{N} \sum_{n=1}^{N} p(x_j, x_n) \gamma_n, \quad 1 \le j \le N., \tag{20}$$

Recall that each datapoint can be represented as a pair $(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, with y_n having a distribution determined by x_n. The sequence γ_n is a set of samples of a random variable defined as

$$\gamma_n := f(y_n) - V(x_n). \tag{21}$$

Equation (8) says that the expected value of the γ variable is $O\left(\Delta t^3\right)$. Equations (18), (19) and (20) are all equivalent definitions of the noise-residual vector.

Table 1. Summary of parameters in experiments. In all these experiments, the number of ...

System	Properties	Figures	ODE parameters	Data samples
Lorenz 63 (26)	Chaotic system in $\mathbb{R}^3$	Figures 3 and 7	$\sigma = 10, \beta = 8/3, \rho = 28$	10^4
Hopf oscillator (27)	Stable periodic cycle in $\mathbb{R}^2$	Figures 4 and 6	$\mu = 1$	10^4
Lorenz 96 (28)	Chaotic system of 10 1-dimensional oscillators, cyclically coupled	Figures 5 and 8	$f = 8.0, N = 10$	2×10^3

Approximation 1. Henceforth, we shall assume a fixed number N of datapoints $\{x_n \ : \ 1 \leq n \leq N\}$ distributed evenly over the domain $\mathcal{X}$, and a subsample $\{\tilde{x}_n \ : \ 1 \leq n \leq N\}$ of size M. This leads to an RKHS subspace

$$\mathcal{H}_M := \text{span}\,\{k(\tilde{x}_n, \cdot) \ : \ 1 \leq n \leq M\}\,.$$

Fix an error level δ_1. Let $B(\mathcal{X})$ be the collection of bounded functions on $\mathcal{X}$. Consider the set

$$\mathcal{H}_M(\delta_1) := \left\{\hat{V} \in B(\mathcal{X}) \ : \ \exists h \in \mathcal{H}_M \text{ s.t. } \left\|\hat{V} - h\right\|_{B(\mathcal{X})} < \delta_1\right\}.$$

Thus, $\mathcal{H}_M(\delta_1)$ contains those bounded functions that can be uniformly approximated within an error limit of δ_1 by an RKHS function in $\mathcal{H}_M$. Also note that

$$\text{clos} \cup_{M \in \mathbb{N}} \mathcal{H}_M(\delta_1) = B(\mathcal{X}), \qquad \forall \delta_1 > 0. \tag{22}$$

We shall assume henceforth that the true image V lies in $\mathcal{H}_M(\delta_1)$. Thus, there is a vector $\hat{a} \in L^2(\nu)$ such that $\|V - K^\nu \hat{a}\| < \delta_1$. This proves Theorem 2 (i). Allowing an error limit of δ_1 we shall assume without loss of generality that $V = K^\nu \bar{a}$.

Approximation 2 Equations (17), (18) and (19) together imply

$$\begin{aligned}\hat{V}_{N,M,\theta} &= K^\nu \left[[K^\nu]^T \left[P^{\beta_X}_{\epsilon_3}\right]^T \left[P^{\beta_X}_{\epsilon_3}\right] [K^\nu] + \theta \right]^{-1} [K^\nu]^T \left[P^{\beta_X}_{\epsilon_3}\right]^T \left[P^{\beta_X} V + \eta_\beta\right] \\ &= K^\nu \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} [K^\nu] \left[B_{N,\epsilon_3} V + \left[P^{\beta_X}_{\epsilon_3}\right]^T \eta_\beta\right].\end{aligned} \tag{23}$$

Incorporating the assumption $V = K^\nu \bar{a}$ gives

$$\hat{V}_{N,M,\theta} = K^\nu \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} \left[A_{N,\epsilon_3,M}\right] \bar{a} + K^\nu \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} [K^\nu] \left[P^{\beta_X}_{\epsilon_3}\right]^T \eta_\beta. \tag{24}$$

Thus, $\hat{V}_{N,M,\theta}$ is the sum of two separate and mutually independent terms, one depending on the true drift and the other only depending on the noise. We shall examine these terms separately.

Approximation 3 Note that the limit in (15) holds irrespective of the nature of μ. When the diffusion is zero, for each $x \in \mathcal{X}$, the conditional measures $\mu_{\mathcal{Y}|x}$ are point measures at zero. In that case according to (18) and (21), $\eta_\nu \equiv 0$. Thus, (15) and (24) combine to give

$$\lim_{N \to \infty, \theta \to 0^+} \left\| \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} \left[A_{N,\epsilon_3,M}\right] \bar{a} - \bar{a} \right\|_{L^2(\nu)} = 0.$$

Since we keep the sub-sampling parameter M fixed and thus ν fixed, K^ν remains a fixed and bounded operator. Therefore, for fixed M, we have:

$$\lim_{N \to \infty, \theta \to 0^+} \left\| K^\nu \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} \left[A_{N,\epsilon_3,M}\right] \bar{a} - K^\nu \bar{a} \right\|_{C(\mathcal{X})} = 0. \tag{25}$$

One important property of the convergence in (25) is that N, θ can be simultaneously taken to their limiting values. This joint convergence is one of the distinguishing features of our technique. The limit in (25) proves Theorem 2 (iii). It remains to prove Theorem 2 (ii). For that, it suffices to bound the error in the second term of (24) by δ_3. In the remainder of the proof our focus will be on this term $K^\nu \left[A_{N,\epsilon_3,M} + \theta\right]^{-1} [K^\nu] \left[P_{\epsilon_3}^{\beta_X}\right]^T \eta_\beta$.

Approximation 4 The self adjoint operator $B_{\beta_X,\epsilon_3} = \left(P_{\epsilon_3}^{\beta_X}\right)^T P_{\epsilon_3}^{\beta_X}$ is also a kernel integral operator with kernel

$$b_{\epsilon_3}^{\beta_X} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}, \quad (z, x') \mapsto \int_{\mathcal{X}} p(z, x') p(x, x') d\beta_X(x)$$

Then according to (19), we have

$$\begin{aligned}
\left[\left(P_{\epsilon_3}^{\beta_X}\right)^T \eta_\beta\right](z) &= \int_{\mathcal{X}} b_{\epsilon_3}^{\beta_X}(z, x') \left[\int_{\mathcal{Y}} f(x', y) d\beta_{\mathcal{Y}|x'}(y) - V(x')\right] d\beta_X(x') \\
&= \int_{\mathcal{X}} b_{\epsilon_3}^{\beta_X}(z, x') \left[\int_{\mathcal{Y}} f(x', y) d\beta_{\mathcal{Y}|x'}(y) - \int_{\mathcal{Y}} f(x', y) d\mu_{\mathcal{Y}|x'}(y)\right] d\beta_X(x') \\
&= \int_{\mathcal{X}} b_{\epsilon_3}^{\beta_X}(z, x') \int_{\mathcal{Y}} \left[f(x', y) - V(x')\right] d\beta_{\mathcal{Y}|x'}(y) d\beta_X(x').
\end{aligned}$$

Equation (21) interprets the noise variable γ to be additive. Thus, we have

$$\left[\left(P_{\epsilon_3}^{\beta_X}\right)^T \eta_\beta\right](z) = \left[\left(P_{\epsilon_3}^{\beta_X}\right) \eta_\beta\right](z) = \frac{1}{N} \sum_{n=1}^{N} b_{\epsilon_3}^{\beta_X}(z, x_n) \gamma_n$$

The distribution of increments of a process defined by an SDE usually lacks an analytic formula. However, due to the continuity of V and G in (1) and the continuity of sample paths of an SDE, we can say that

$$\lim_{\Delta t \to 0^+} \text{var}(\gamma) = \mathbb{E}_{\mu_{stat}} G G^T.$$

In other words, for positive Δt, γ converges in distribution to a Gaussian variable on $\mathbb{R}^d$ whose variance matrix is a strictly positive definite matrix. Now by the central limit theorem we have

$$X(N, z, \epsilon_3) := N^{1/2} \left[\left(P_{\epsilon_3}^{\beta_X}\right) \eta_\beta\right](z) \xrightarrow[N \to \infty]{a.s.} \mathcal{N}\left(0, \text{var}(\gamma)\right) + O\left(\Delta t^3\right).$$

We have declared the quantity to be a random variable $X(N, z, \epsilon_3)$ to denote its dependence on N, bandwidth ϵ_3 and initial location z. Let us fix an error bound δ_3. Thus, by the multivariate Chebyshev inequality[36] we have

$$\text{Prob}\left\{\left|P_{\epsilon_3}^{\beta_X} \eta_\beta - O\left(\Delta t^3\right)\right| < \delta_3\right\} > 1 - O\left(\frac{d}{N^{1/2} \delta_3^2}\right).$$

So, to bound the uncertainty by δ_4, we would need $N = O\left(\frac{d^2}{\delta_3^4 \delta_4^2}\right)$, as claimed. This completes the proof of Theorem 2.

This concludes the description of our methods and the underlying theory. We next describe some numerical experiments to test our technique on.

4. EXAMPLES

In this section, we put our algorithms to the test. The systems that we investigate are as follows:

Table 2. Parameters which remain constant in the experiments

Variable	η_3	(I, w)	θ
Interpretation	Parameter for selecting the bandwidth ϵ_3 of the Markov kernel, based on Algorithm 2	Dimensions of the image	Ridge regression parameter, as described in (11)
Value	4	$(1, 1)$	$0.01 \times \|K^\nu\|$

Table 3. Auto-tuned parameters in experiments. The strategy outlined in (32) is used to determine the three bandwidth parameters $\epsilon_1, \epsilon_2, \epsilon_3$

Ridge regression parameter ϵ	Bandwidth ϵ_1 of Gaussian kernel $\mathcal{G}$	Bandwidth ϵ_2 of RKHS kernel k	Bandwidth ϵ_3 of Markov kernel p
0.1	Chosen to achieve sparsity 0.3%	Chosen to achieve sparsity 1%	Chosen to achieve sparsity 1%

1. **Lorenz 63.** This is the benchmark problem for studying chaotic phenomena in three dimensions, which is the lowest dimension in which chaos is possible[37].

$$\begin{aligned} \frac{d}{dt}x_1(t) &= \sigma(x_2 - x_1) \\ \frac{d}{dt}x_2(t) &= x_!(\rho - x_3) - x_2 \\ \frac{d}{dt}x_3(t) &= x_1 x_2 - \beta x_3 \end{aligned} \tag{26}$$

2. **Hopf oscillator.** The Hopf oscillator is used as a simple parametric model to study the phenomenon of Hopf bifurcations. Given a constant p, the ODE is[38]

$$\begin{aligned} \frac{d}{dt}x_1(t) &= -x_2 + x_1\left(p - \left(x_1^2 + x_2^2\right)\right) \\ \frac{d}{dt}x_2(t) &= x_1 + x_2\left(p - \left(x_1^2 + x_2^2\right)\right) \end{aligned} \tag{27}$$

3. **Lorenz 96.** The Lorenz 96 model is a dynamical system formulated by Edward Lorenz in 1996

$$\frac{d}{dt}x_n(t) = (x_{n+1} - x_{n-2})\,x_{n-1} - x_n + F, \qquad \forall n \in 1, \ldots, N. \tag{28}$$

This model mimics the time evolution of an unspecified weather measurement collected at N equidistant grid points along a latitude circle of the earth[39]. The constant F is known as the *forcing* function. The system (28) is benchmark problem in data assimilation.

The steps in each experiment were as follows:

1. The parameters for the ODEs were set according to Table 1.
2. The ODEs were converted into an SDE with the help of a diagonal diffusion term

$$G(x)_{i,i} = \sigma V(x)_i, \qquad i = 1, \ldots, d. \tag{29}$$

3. The constant σ, which controls the level of noise, was varied among the values $\{0.1, 0.2, 0.5\}$ respectively.
4. Algorithm 2 was applied to a sample path of the SDE. The parameters for the algorithm were auto-tuned as described in Table 2.
5. The choice of kernel k was a diffusion kernel as in (30), and a Gaussian Markov normalized kernel as in (10).
6. Some of the algorithmic parameters were auto-tuned according to the hyper-parameters listed in Table 3.

The choice of noise levels does not follow any specific design, other than providing an increasing sequence. Recall that a higher noise level requires a larger number of data points to approximate the conditional expectation. The highest value of 0.5 was chosen heuristically based on the size $\approx 10^4$ of the dataset being used. We next describe our parameter choices, and methodology for evaluating our results.

Table 4. Relative L^2 error in experiments

Experiment	Noise 0.1	Noise 0.2	Noise 0.5
Hopf oscillator (27)	0.02	0.03	0.06
Lorenz 63 (26)	0.08	0.1	0.14
Lorenz 96 (28) - 5 cells	0.31	0.27	0.265

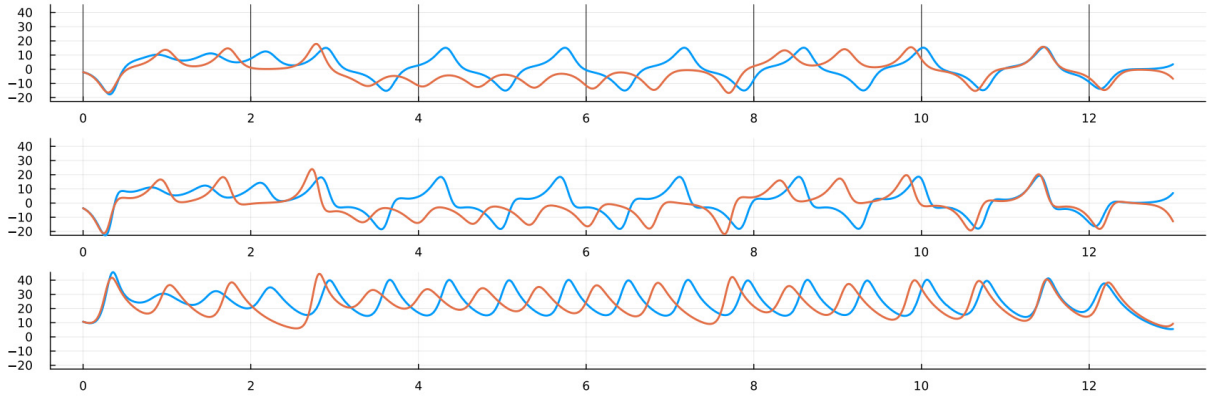

Figure 7. Reconstruction of the Lorenz 63 system. The trajectories corresponding to the true and reconstructed drifts are in blue and orange, respectively.

Choice of kernel Our choice of kernel in all the experiments is the *diffusion* kernel[40,41]. There are many versions of the diffusion kernel. We choose the following:

$$\begin{aligned} k^{\mu}_{\mathrm{diff},\epsilon}(x,y) &= \frac{k_{\mathrm{Gauss},\epsilon}(x,y)}{\deg_l(x)\deg_r(y)}, \\ \deg_r(x) &:= \int_X k_{\mathrm{Gauss},\epsilon}(x,y)d\mu(y), \qquad \deg_l(x) := \int_X k_{\mathrm{Gauss},\epsilon}(x,y)\frac{1}{\deg_r(x)}d\mu(y). \end{aligned} \tag{30}$$

Diffusion kernels have been shown to be good approximants of the local geometry in various situations[42–44], and are a natural choice for nonparametric learning. We go one step further and perform a symmetrization:

$$\rho(x)k^{\mu}_{\mathrm{diff},\epsilon}(x,y)\rho(y)^{-1} = \tilde{k}^{\mu}_{\mathrm{diff},\epsilon}(x,y) = \frac{k_{\mathrm{Gauss},\epsilon}(x,y)}{\left[\deg_r(x)\deg_r(y)\deg_l(x)\deg_l(y)\right]^{1/2}}, \tag{31}$$

where

$$\rho(z) = \deg_l(z)^{1/2}/\deg_r(z)^{1/2}.$$

The kernel $\tilde{k}^{\mu}_{\mathrm{diff},\epsilon}$ from (31) is clearly symmetric. Since it is built from the s.p.d. kernel $k_{\mathrm{Gauss},\epsilon}$, $\tilde{k}^{\mu}_{\mathrm{diff},\epsilon}$ is s.p.d. too and thus generates an RKHS of its own. Moreover, the kernel $k^{\mu}_{\mathrm{diff},\epsilon}$ can be symmetrized using a degree function ρ, which is bounded as well as bounded away from 0. Such a kernel will be called *RKHS-like*. Let M_ρ be the multiplication operator with ρ. Then

$$\operatorname{ran} K^{\mu}_{\mathrm{diff},\epsilon} = \operatorname{ran} M_\rho \circ \tilde{K}^{\mu}_{\mathrm{diff},\epsilon}.$$

Again, because of the properties of ρ, both M_ρ and its inverse are bounded operators. Thus, there is a bijection between the RKHS generated by $\tilde{k}^{\mu}_{\mathrm{diff},\epsilon}$, and the range of the integral operator $K^{\mu}_{\mathrm{diff},\epsilon}$.

Bandwidth selection The proposed algorithm depends on several choices of bandwidth parameters, which also play the role of a smoothing parameter. We now present our method of selecting a bandwidth ϵ for constructing a kernel over a dataset $\mathcal{D}$. It depends on the choice of a threshold $\eta \in (0,1)$, and a zero threshold $\theta_{zero} = 10^{-14}$.

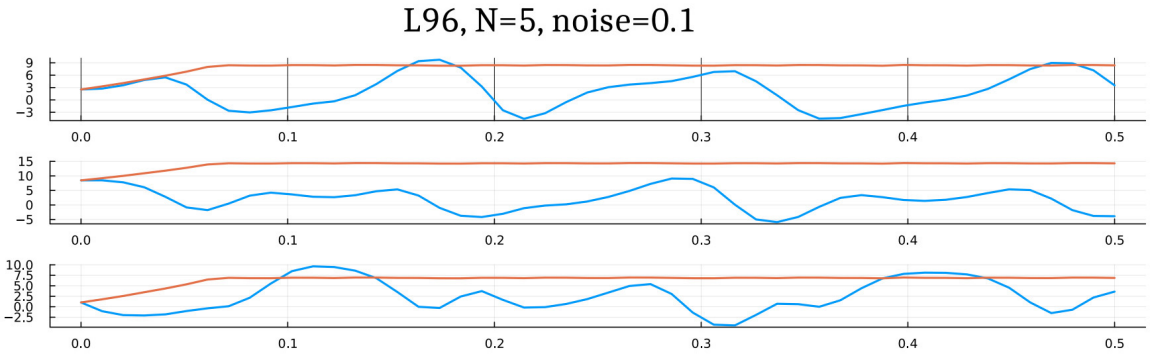

Figure 8. Reconstruction of the Lorenz 96 system with 5 cells. The orbits of the true and reconstructed systems diverge rapidly due to the presence of positive Lyapunov exponents in the L96 system. See Figure 9 for a closer look at the plots. The trajectories corresponding to the true and reconstructed drifts are in blue and orange, respectively.

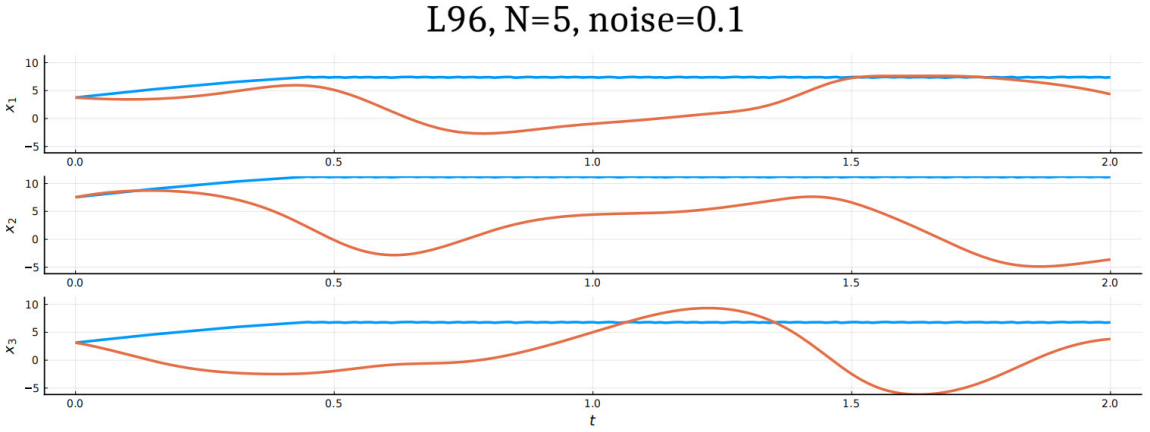

Figure 9. Reconstruction of the Lorenz 96 system with 5 cells. This figure offers a closer look at the graphs of Figure 8. The trajectories corresponding to the true and reconstructed drifts are in blue and orange, respectively.

We first choose a subsample $\mathcal{D}'$ of the data set $\mathcal{D}$ and construct the set $\mathcal{S}$ of all possible pairwise squared distances. Usually, it suffices to choose $\mathcal{D}'$ to be 0.1 fraction of the dataset $\mathcal{D}$, equidistributed throughout $\mathcal{D}$. Given the threshold η, we set ϵ such that η-fraction of the numbers in $\mathcal{S}$ are greater than $\epsilon\theta$, where θ is the threshold:

$$\theta = -1/\ln(\theta_{zero})\,. \tag{32}$$

Thus ϵ is chosen so that an η fraction of the values in the set $\left\{e^{-\|x-x'\|^2/\epsilon} \;:\; x, x' \in \mathcal{D}'\right\}$ are greater than θ_{zero}. The strategy outlined in (32) is used to determine the three bandwidth parameters $\epsilon_1, \epsilon_2, \epsilon_3$ in Table 2.

Evaluating the results There is no certain way of evaluating the efficacy of a reconstruction technique for chaotic systems, as argued in [45,46]. In any data-driven approach such as ours, the reconstruction is done in dimensions higher than the original system. As a result, new directions of instability are introduced, and even if the drift is reconstructed with good precision, the simulation results might be completely different. For that purpose, our primary means of evaluating the accuracy of our drift reconstruction is by a pointwise evaluation of the reconstruction error, as shown in Figures 3, 4 and 5. The average errors are tabulated in Table 4. Also see Figures 6, 7, 8 and 9 for a comparison of orbits from the true system and the reconstructed system. It can be noted in the last two figures that the orbits of the reconstructed system get stuck in a false fixed point. This is an artifact of the technique of making out of sample evaluations, which we discuss next.

Evaluating kernel functions Any kernel function $\phi(x) = \sum_{n=1}^{N} a_n \tilde{K}^{\mu}_{\text{diff},\epsilon}(x, x_n)$ created from a diffusion kernel (30) can be evaluated at an arbitrary point x in $\mathbb{R}^d$ as

$$\phi(x) = \frac{1}{N} \left[\vec{\rho}_l(x)\vec{\rho}_r(x)\right]^{-0.5} \sum_{n=1}^{N} k_{Gauss,\epsilon}(x, x_n) w_n,$$

where the other terms are given by

$$w_n = a_n \left[\deg_r(x_n) \deg_l(x_n)\right]^{-0.5}, \ \vec{\rho}_r(x) = \frac{1}{N} \sum_{n=1}^{N} k_{Gauss,\epsilon}(x, x_n), \ \vec{\rho}_l(x) = \frac{1}{N} \sum_{n=1}^{N} k_{Gauss,\epsilon}(x, x_n)/\deg_r(x_n).$$

Note that the vector $\vec{w} = (w_1, \ldots, w_N)$ is independent of x and is computed during the creation of the kernel matrix. While the computation is straightforward, a numerical issue arises when the test point x is far from the data set $\{x_n \ : \ n \in 1, \ldots, N\}$. In that case, the vector $\left(k_{Gauss,\epsilon}(x, x_n)\right)_{n=1}^{N}$ has all entries below the machine precision. Therefore, the constant $\vec{\rho}_r(x)$ is also below machine precision. The occurrence of such near-zero numbers could lead to several computational anomalies. We use a nearest point approximation for evaluating the kernel sum in such cases. This also leads to false fixed points emerging at locations far from the dataset. We discuss this issue more in our concluding section.

Summary of results Figure 2 represents some typical SDE orbits that were provided as input to the algorithms. The results of the reconstruction have been displayed in Figures 3, 4, and 5 respectively. The expected trend displayed is that data created out of higher noise levels lead to larger errors in reconstruction. While the plots present the pointwise comparisons, the root mean squared errors are summarized in Table 4. The trend can also be explained by the convergence analysis of the conditional expectation scheme.

An important test for drift reconstruction algorithms is their efficacy in reconstructing trajectories. To this end, we have compared the trajectories generated by the true drift, with trajectories generated by the simulated drift. The trajectories show excellent match for the periodic Hopf oscillator [Figure 6], and a close match for the Lorenz 63 model [Figure 7]. The Lorenz-63 model is a chaotic system, and the presence of positive Lyapunov exponents leads to an exponential divergence of trajectories. The gradual divergence of the trajectories in Figure 7 is indicative of a good reconstruction. It is also notable that the reconstructed orbit, even though divergent, remains topologically close to the true attractor X. The true Lorenz attractor is an isolated invariant set, with an attracting neighborhood. The topological proximity of the reconstruction indicates that the reconstructed drift preserves the topological confinement of a neighborhood of X.

The 5-dimensional Lorenz-96 model presents a much steeper challenge, as evident from the trajectory comparison plots in Figures 8 and 9 respectively. As explained in Section and Theorem 2, the rate of convergence depends on the number of data samples N and the standard deviation σ of the noise. The higher the dimension d of the phase-space, the higher σ will be. This means a larger number of data samples are required to accurately sample the phase space. This is the well known and inevitable curse of dimensionality, given a fixed error bound δ, the number of samples required to accurately sample the distribution scales as $\delta^{1/N}$. This practical challenge can also be interpreted as the case of under-sampling in high dimensions. This problem has been tackled in [47], but only for the case of *overparameterized* SDEs. We end this article by proposing a technique that can overcome this challenge on-parametrically, for certain drifts having redundancy in their structure.

5. SPARSITY

The mechanism by which Algorithm 1 approximates the conditional expectations can be broadly described as computing averages w.r.t. various empirical measures. Empirical measure is the measure directly available

to the algorithm. It is the discrete probability measure whose supports are the data points. The key principle behind the technique is that the empirical measure provides a weak approximation of the unknown measure μ referred to in Lemma 0.1. In our next assumption, we make a precise statement on how each component of the drift is bound to no more than m other coordinates. This number m is much smaller compared to the dimension d. Moreover, the functional dependence is also limited in type, as described formally below:

Assumption 3 *Given the SDE* (1)*, there is*

(i) an integer $m << d$ *called the* dependency dimension*;*
(ii) a finite collection of functions $\tilde{V}^l : \mathbb{R}^m \to \mathbb{R}$*, for* $1 \leq l \leq L$*;*
(iii) and for each $i \in \{1, \ldots, d\}$ *a collection Input*(i) *of m coordinates* $\left(k_{i,1}, \ldots, k_{i,m}\right) \subset \{1, \ldots, d\}$ *and an index* $\ell(i) \in 1, \ldots L$*;*

such that

$$V(x)_i = \tilde{V}^{\ell(i)}\left(x_{k_{i,1}}, \ldots, x_{k_{i,m}}\right), \quad \forall 1 \leq i \leq d. \tag{33}$$

Assumption 3 says that effectively, each component of the drift V is a repeated instance of a lower dimensional functional unit $\tilde{V}^{(l)}$. Such drifts occur in many natural complex phenomena, such as natural complex dynamical systems[48], networks of low dimensional dynamical systems[49] and sub-grid modeling[50,51]. We now derive the precise nature of this expectation. For each $1 \leq i \leq d$, define projection maps

$$\pi^i : \mathbb{R}^d \to \mathbb{R}^m, \quad x \mapsto \left(x_{k_{i,1}}, \ldots, x_{k_{i,m}}\right).$$

These projection maps create the following commutation with the coordinate projections $\left\{\text{proj}_i \ : \ 1 \leq i \leq d\right\}$, as given below:

$$\begin{array}{ccc} \mathbb{R}^d & \xrightarrow{\pi^i} & \mathbb{R}^m \\ {\scriptstyle V}\downarrow & & \downarrow{\scriptstyle \tilde{V}^{\ell(i)}} \\ \mathbb{R}^d & \xrightarrow[\text{proj}_i]{} & \mathbb{R} \end{array}, \quad \forall 1 \leq i \leq d. \tag{34}$$

Taking $\phi = \text{proj}_i$, we get the following equality with the aid of the limiting relation (3) and commutation (34):

$$\frac{d}{dt}\mathbb{E}\left(X_i(t, x_0)|x_0\right) = \left(\text{proj}_i V\right)(x_0) = \tilde{V}^{\ell(i)} \circ \pi^i(x_0), \quad \forall 1 \leq i \leq d.$$

Recall the function f from (7). Then, similar to (8), we can write

$$V_i = \tilde{V}^{\ell(i)} = \mathbb{E}_\mu\left(\text{proj}_i \circ f \,|\, \pi_i\right), \quad \forall 1 \leq i \leq d. \tag{35}$$

Equation (35) says that each component of the drift $\tilde{V}$ can be recovered as a conditional expectation involving measurements of only a part of the coordinates of the full d-dimensional state space. It is also important to note that this equality holds for each i. This means that (35) provides two advantages from a data-handling perspective. Firstly, even if d is large, one does not need simultaneous measurements of each of the d coordinates at every time step. Secondly, the number of coordinates or variables involved in the calculation is m, which could be much smaller than d.

To correctly utilize the strength of (35), we impose some restrictions on the data being used. A *causally complete snapshot* corresponding to an initial condition x_0 and a coordinate $i \in 1, \ldots, d$ is an $(4+m)$-vector:

$$\left(X^{(i)}(3\Delta t, x_0), X^{(i)}(2\Delta t, x_0), X^{(i)}(\Delta t, x_0), X^{(i)}(t, x_0), \pi^i(x_0)\right),$$

Each initial condition x_0 and choice of coordinate i will lead to a different instantiation of this vector. The initial conditions need not be selected from a single path; they only need to be equidistributed w.r.t. μ_{stat}.

The conditional expectation scheme (35) can be numerically realized in a manner exactly analogous to Algorithm 2. The convergence rates undergo an analysis similar to Section . The scheme in (35) allows the components of the drift to be mined from high dimensional data, in which trajectories may be short, and partially missing in data. A numerical implementation of (35) to real-world high dimensional data is an important and promising project. These numerical investigations will be conducted in our forthcoming work.

6. CONCLUSIONS

Thus, we have demonstrated using theory and numerical examples that the drift component of an SDE can be extracted from data, using principles from Probability theory and the theory of Kernel integral operators. The drift can be expressed as a conditional expectation involving random variables, on a certain probability space. This conditional expectation in turn can be estimated as a linear regression problem using techniques from kernel integral operators. The algorithmic procedure that is created is completely data-driven. It receives as inputs snapshots of trajectories of the unknown SDE, and the only requirement is that the initial condition in each snapshot is equidistributed w.r.t. some stationary measure of the SDE. The numerical method worked well for the chaotic and quasiperiodic systems of dimensions less than 6 that we investigated. The performance deteriorates as the dimension increases.

Choice of parameters The three main tuning parameters in our algorithm are the Ridge regression θ and kernel bandwidth parameters $\epsilon_1, \epsilon_2, \epsilon_3$. Among these, the most important is the smoothing parameter ϵ_3 which determines the effective radius of integration. The smaller ϵ_3 is, the higher the number N of total samples has to be. Thus, a larger smoothing radius ϵ_3 leads to a faster convergence with N. On the other hand, a larger ϵ_3 and N also increases the condition number of the Markov matrix $[P]$, which could adversely affect the accuracy of the solution to the linear least squares problem (13). There is no recipe for tuning these parameters that would work uniformly well in all applications.

Choice of kernel Note that Algorithm 1 does not specify the kernel, and any RKHS-like kernel such as the diffusion kernel would be sufficient. The convergence laid down in Lemma 0.1 also does not depend on the choice of kernel. However, when finite data is used, the choice of kernel becomes important. The question of an optimal kernel is important and unsolved in machine learning[52–54]. Our framework for estimating drift provides a more objective setting for this question.

Out-of-sample extensions The technique that we presented provides a robust estimation technique for the drift. The reason it is well suited to a general nonlinear and data-driven technique is the use of kernel methods. The downside of kernel methods is the difficulty in extending the function beyond the compactly supported dataset. Thus, our technique provides a good phase-portrait, but it may not be reliable as a simulator or alternative model to the true system if the latter is chaotic and high dimensional. The issue of out of sample extensions[55,56] and topological containment[57,58] of simulation models is separate from the preliminary task of estimating the drift. This is yet another important area of development.

The present article reconstructs the drift as a sum of kernel sections, i.e., as a function. Kernel methods are just one of the many ways of representing a function. The next goal is to obtain a different representation that is more suitable for retaining the topology, but still aims to achieve the identity (8). A recent work by the author[59] presents a reconstruction strategy that preserves the topology of the targeted attractor, at the cost of replacing the deterministic dynamics law by a Markov process with arbitrarily low noise. The continuous time analog of this strategy is the idea of combinatorial drifts by Mrozek *et al.*[60,61]. Another promising technique directed towards preserving the attractor topology is the matrix-based approach[62]. A fourth option is the recent idea of representing vector fields as 1-forms[63]. A combination of any of these topologically motivated techniques, with the statistical methods presented in this article, offers an interesting avenue for further

research.

DECLARATIONS

Acknowledgments
The author is grateful to the referees and editors for the detailed comments and reflections on the paper. They helped to improve the paper significantly.

Authors' contributions
The author contributed solely to the article.

Availability of data and materials
All the raw data and experimental datasets are sample trajectories of SDEs. The trajectories are simulated using the *SDEProblem* feature of DifferentialEquations.jl package in Julia language. The integration technique was the Euler-Maruyama method. The data may be regenerated using the parameters specified in Table 1.

Financial support and sponsorship
None.

Conflicts of interest
The author declared that there are no conflicts of interest.

Ethical approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

REFERENCES

1. Sun, J.; Wang, H. Stability analysis for highly nonlinear switched stochastic systems with time-varying delays. *Complex. Eng. Syst.* **2022**, *2*, 17 DOI
2. Song, K.; Bonnin, M.; Traversa, F. L.; Bonani, F. Stochastic analysis of a bistable piezoelectric energy harvester with a matched electrical load. *Nonlinear. Dyn.* **2023**, *111*, 16991-7005. DOI
3. Zhang, Y.; Duan, J.; Jin, Y.; Li, Y. Discovering governing equation from data for multi-stable energy harvester under white noise. *Nonlinear. Dyn.* **2021**, *106*, 2829-40. DOI
4. Neves, W.; Olivera, C. Stochastic transport equations with unbounded divergence. *J. Nonlinear. Sci.* **2022**, *32*, 60. DOI
5. Espanol, P. Fluid particle model. *Phys. Rev. E.* **1998**, *57*, 2930. DOI
6. Gay-Balmaz, F.; Holm, D. Stochastic geometric models with non-stationary spatial correlations in Lagrangian fluid flows. *J. Nonlinear. Sci.* **2018**, *28*, 873-904. DOI
7. Arnold, L. Stochastic differential equations : theory and applications. Wiley Interscience; 1974. Available from: https://cir.nii.ac.jp/crid/1970867909896889156 [Last accessed on 24 Oct 2025].
8. Philipp, F.; Schaller, M.; Worthmann, K.; Peitz, S.; Nüske, F. Error bounds for kernel-based approximations of the Koopman operator; 2023. DOI.
9. Doob, J. L. Stochastic processes. Wiley; 1953. Available from: https://www.wiley.com/en-us/Stochastic+Processes-p-9780471523697[Last accessed on 24 Oct 2025].
10. Huang, W.; Ji, M.; Liu, Z.; Yi, Y. Integral identity and measure estimates for stationary Fokker-Planck equations. *Ann. Probab.* **2015**, *43*, 1712-30. DOI
11. Harris, T. The existence of stationary measures for certain Markov processes. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability. 1956; pp. 113-24. Available from: https://apps.dtic.mil/sti/html/tr/AD0604937/[Last accessed on 24 Oct 2025].

12. Huang, W.; Ji, M.; Liu, Z.; Yi, Y. Concentration and limit behaviors of stationary measures. *Phys. D.* **2018**, *369*, 1-17. DOI
13. Shalizi, C.; Kontorovich, A. Almost none of the theory of stochastic processes. Lecture Notes; 2010. Available from: www.stat.cmu.edu/~cshalizi/almost-none/ [Last accessed on 24 Oct 2025].
14. Stanton, R. A nonparametric model of term structure dynamics and the market price of interest rate risk. *J. Finance.* **1997**, *52*, 1973-2002. DOI
15. Gouesbet, G. Reconstruction of the vector fields of continuous dynamical systems from numerical scalar time series. *Phys. Rev. A.* **1991**, *43*, 5321. DOI
16. Tsutsumi, N.; Nakai, K.; Saiki, Y. Constructing differential equations using only a scalar time-series about continuous time chaotic dynamics. *Chaos.* **2022**, *32*, 091101. DOI
17. Gouesbet, G.; Letellier, C. Global vector-field reconstruction by using a multivariate polynomial L 2 approximation on nets. *Phys. Rev. E.* **1994**, *49*, 4955. DOI
18. Cao, J.; Wang, L.; Xu, J. Robust estimation for ordinary differential equation models. *Biometrics.* **2011**, *67*, 1305-13. DOI
19. Aït-Sahalia, Y.; Hansen, L. P.; Scheinkman, J. A. Chapter 1 - Operator methods for continuous-time Markov processes. In: Handbook of financial econometrics: tools and techniques. Elsevier; 2010. pp. 1-66. DOI
20. Garcia, C.; Otero, A.; Félix, P.; Presedo, J.; Márquez, D. G. Nonparametric estimation of stochastic differential equations with sparse Gaussian processes. *Phys. Rev. E.* **2017**, *96*, 022104. DOI
21. Devlin, J.; Husmeier, D.; Mackenzie, J. Optimal estimation of drift and diffusion coefficients in the presence of static localization error. *Phys. Rev. E.* **2019**, *100*, 022134. DOI
22. Darcy, M.; Hamzi, B.; Livieri, G.; Owhadi, H.; Tavallali, P. One-shot learning of stochastic differential equations with data adapted kernels. *Phys. D.* **2023**, *444*, 133583. DOI
23. Batz, P.; Ruttor, A.; Opper, M. Approximate Bayes learning of stochastic differential equations. *Phys. Rev. E.* **2018**, *98*, 022109. DOI
24. Ruttor, A.; Batz, P.; Opper, M. Approximate Gaussian process inference for the drift of stochastic differential equations. 2013; Available from: https://proceedings.neurips.cc/paper_files/paper/2013/file/021bbc7ee20b71134d53e20206bd6feb-Paper.pdf [Last accessed on 24 Oct 2025].
25. Ella-Mintsa, E. Nonparametric estimation of the diffusion coefficient from i.i.d. S.D.E. paths. *Stat. Inference. Stoch. Process.* **2024**, *27*, 585. DOI
26. Davis, W.; Buffett, B. Estimation of drift and diffusion functions from unevenly sampled time-series data. *Phys. Rev. E.* **2022**, *106*, 014140. DOI
27. Ye, F.; Yang, S.; Maggioni, M. Nonlinear model reduction for slow-fast stochastic systems near unknown invariant manifolds. *J. Nonlinear. Sci.* **2024**, *34*, 22. DOI
28. Wang, Z.; Ma, Q.; Yao, Z.; Ding, X. The magnus expansion for stochastic differential equations. *J. Nonlinear. Sci.* **2020**, *30*, 419-47. DOI
29. Yin, X.; Zhang, Q. Backstepping-based state estimation for a class of stochastic nonlinear systems. *Complex. Eng. Syst.* **2022**, *2*, 1. DOI
30. Das, S. Conditional expectation using compactification operators. *Appl. Comput. Harmon. Anal.* **2024**, *71*, 101638. DOI
31. Das, S.; Giannakis, D. Delay-coordinate maps and the spectra of Koopman operators. *J. Stat. Phys.* **2019**, *175*, 1107-45. DOI
32. Das, S.; Giannakis, D. Koopman spectra in reproducing kernel Hilbert spaces. *Appl. Comput. Harmon. Anal.* **2020**, *49*, 573-607. DOI
33. Das, S.; Mustavee, S.; Agarwal, S. Data-driven discovery of quasiperiodically driven dynamics. *Nonlinear. Dyn.* **2025**, *113*, 4097–120. DOI
34. Das, S.; Mustavee, S.; Agarwal, S.; Hassan, S. Koopman-theoretic modeling of quasiperiodically driven systems: example of signalized traffic corridor. *IEEE Trans. Syst. Man. Cyber. Syst.* **2023**, *53*, 4466-76. DOI
35. Giannakis, D.; Das, S.; Slawinska, J. Reproducing kernel Hilbert space compactification of unitary evolution groups. *Appl. Comput. Harmon. Anal.* **2021**, *54*, 75-136. DOI
36. Marshall, A.; Olkin, I. Multivariate chebyshev inequalities. *Anna. Math. Stat.* **1960**, *31*, 1001-14. DOI
37. Lorenz, E. Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **1969**, *26*, 636-46. DOI
38. Strogatz, S. *Nonlinear dynamics and chaos with applications to physics, biology, chemistry, and engineering*, 2nd ed.; CRC Press, 2015. DOI
39. Lorenz, E.; Emanuel, K. Optimal sites for supplementary weather observations: simulation with a small model. *J. Atmos. Sci.* **1998**, *55*, 399-414. DOI
40. Marshall, N.; Coifman, R. Manifold learning with bi-stochastic kernels. *IMA. J. Appl. Math.* **2019**, *84*, 455-82. DOI
41. Wormell, C. L.; Reich, S. Spectral convergence of diffusion maps: Improved error bounds and an alternative normalization. *SIAM. J. Numer. Anal.* **2021**, *59*, 1687-734. DOI
42. Coifman, R.; Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **2006**, *21*, 5-30. DOI
43. Hein, M.; Audibert, J. Y.; von Luxburg , U. From graphs to manifolds-weak and strong pointwise consistency of graph Laplacians. In: International Conference on Computational Learning Theory. Springer; 2005. pp. 470-85. DOI
44. Vaughn, R.; Berry, T.; Antil, H. Diffusion maps for embedded manifolds with boundary with applications to PDEs. *Appl. Comput. Harmonic. Anal.* **2024**, *68*, 101593. DOI
45. Berry, T.; Das, S. Learning theory for dynamical systems. *SIAM. J. Appl. Dyn.* **2023**, *22*, 2082-122. DOI
46. Berry, T.; Das, S. Limits of learning dynamical systems. *SIAM. Rev.* **2025**, *67*, 107-37. DOI
47. Wang, S.; Blanchet, J.; Glynn, P. An efficient high-dimensional gradient estimator for stochastic differential equations. *Adv. Neural. Inf. Proc. Syst.* **2024**, *37*, 88045. Available from: https://proceedings.neurips.cc/paper_files/paper/2024/hash/a0cd56b91305239e2580dd9440b2e155-Abstract-Conference.html [Last accessed on 24 Oct 2025].
48. Chen, D.; Chen, J.; Zhang, X.; et al. Critical nodes identification in complex networks: a survey. *Complex. Eng. Syst.* **2025**, *5*, 11. DOI

49. Yang, C.; Suh, C. On controlling dynamic complex networks. *Phys. D.* **2022**, *441*, 133499. DOI
50. Guan, Y.; Subel, A.; Chattopadhyay, A.; Hassanzadeh, P. Learning physics-constrained subgrid-scale closures in the small-data regime for stable and accurate LES. *Phys. D.* **2023**, *443*, 133568. DOI
51. Boutros, D.; Titi, E. Onsager's conjecture for subgrid scale α-models of turbulence. *Phys. D.* **2023**, *443*, 133553. DOI
52. Narayan, A.; Yan, L.; Zhou, T. Optimal design for kernel interpolation: applications to uncertainty quantification. *J. Comput. Phys.* **2021**, *430*, 110094. DOI
53. Baraniuk, R.; Jones, D. A signal-dependent time-frequency representation: optimal kernel design. *IEEE. Trans. Signal. Process.* **1993**, *41*, 1589-602. DOI
54. Crammer, K.; Keshet, J.; Singer, Y. Kernel design using boosting. In Part of Advances in Neural Information Processing Systems 15 (NIPS 2002); 2002. Available from: https://papers.nips.cc/paper_files/paper/2002/hash/dd28e50635038e9cf3a648c2dd17ad0a-Abstract.html [Last accessed on 24 Oct 2025].
55. Vural, E.; Guillemot, C. Out-of-sample generalizations for supervised manifold learning for classification. *IEEE. Trans. Image. Process.* **2016**, *25*, 1410-24. DOI
56. Pan, B.; Chen, W. S.; Chen, B.; Xu, C.; Lai, J. Out-of-sample extensions for non-parametric kernel methods. *IEEE. Trans. Neural. Netw. Learn. Syst.* **2016**, *28*, 334-45. DOI
57. Kaczynski, T.; Mrozek, M.; Wanner, T. Towards aformal tie between combinatorial and classical vector field dynamics. *J. Comput. Dyn.* **2016**, *3*, 17-50. DOI
58. Mrozek, M.; Wanner, T. Creating semiflows on simplicial complexes from combinatorial vector fields. *J. Dif. Eq.* **2021**, *304*, 375-434. DOI
59. Das, S. Reconstructing dynamical systems as zero-noise limits; 2024. DOI
60. Mrozek, M. Topological invariants, mulitvalued maps and computer assisted proofs in dynamics. *Comput. Math. Appl.* **1996**, *32*, 83-104. DOI
61. Mischaikow, K.; Mrozek, M. Isolating neighborhoods and chaos. *Japan. J. Ind. Appl. Math.* **1995**, *12*, 205-36. DOI
62. Du, Q.; Yang, H. Computation of robust positively invariant set based on direct data-driven approach. *Complex. Eng. Syst.* **2024**, *4*, 24. DOI
63. Das, S.; Giannakis, D.; Gu, Y.; Slawinska, J. Learning dynamical systems with the spectral exterior calculus; 2025. DOI