SIAM J. Applied Dynamical Systems
Vol. 23, No. 2, pp. 1504–1539

On Higher Order Drift and Diffusion Estimates for Stochastic SINDy*

Mathias Wanner† and Igor Mezić†

Abstract. The sparse identification of nonlinear dynamics (SINDy) algorithm can be applied to stochastic differential equations (SDEs) to estimate the drift and the diffusion function using data from a realization of the SDE. The SINDy algorithm requires sample data from each of these functions, which is typically estimated numerically from the data of the state. We analyze the performance of the previously proposed estimates for the drift and the diffusion function to give bounds on the error for finite data. However, since this algorithm only converges as both the sampling frequency and the length of trajectory go to infinity, obtaining approximations within a certain tolerance may be infeasible. To combat this, we develop estimates with higher orders of accuracy for use in the SINDy framework. For a given sampling frequency, these estimates give more accurate approximations of the drift and diffusion functions, making SINDy a far more feasible system identification method.

Key words. stochastic differential equations, system identification, numerical methods, SINDy

MSC codes. 37H99, 37M15, 60H35, 65C40, 93E12

DOI. 10.1137/23M1567011

1. Introduction.

For many dynamical systems, data might be abundant while there remain no analytic models to describe the system. These systems may be too complex, may have too large a dimension, or may be too poorly understood to model using first principles. For these reasons, data-driven modeling has become important for applications in science and engineering. There are a wide variety of system identification methods, ranging from classical methods [19] to dynamic mode decomposition and Koopman operator methods [29, 36, 23, 35] to neural networks [25, 18] and many others. These methods vary in their complexity, training methods, model sizes, and interpretability. Sparse identification of nonlinear dynamics (SINDy) is a method which allows for some complexity (allowing nonlinear models over only linear ones), while the sparse solution promotes simple, interpretable models.

The SINDy algorithm, developed by Brunton, Proctor, and Kutz [3] estimates the parameters of an ordinary differential equation (ODE) from data. It does this by using a dictionary of functions and finding a sparse representation of the derivative in this dictionary. The data for the derivative can be obtained using finite differences of data from the state. For ODEs, the performance of this algorithm has been analyzed in [37].

*Received by the editors April 27, 2023; accepted for publication (in revised form) by G. Gottwald February 24, 2024; published electronically June 14, 2024.
https://doi.org/10.1137/23M1567011
Funding: This research was supported by grants ARO-MURI W911NF-17-1-0306 and NSF EFRI C3 SoRo 1935327.
†Department of Mechanical Engineering, University of California, Santa Barbara, Santa Barbara, CA 93106 USA (mwanner@ucsb.edu, mezic@ucsb.edu).

SINDy has several extensions and adaptations; it has also been extended to identify control systems [4, 14], adapted to systems with implicit solutions [20, 13], and formulated in ways to improve its robustness to noise [9, 22, 21], to name a few. Additionally, different methods for computing the sparse solution have been proposed, including LASSO [33], the sequential thresholding presented in the original paper [3].

The problem of system identification can be similarly posed for stochastic differential equations (SDEs). Many systems, due to their complexity, separation of timescales, or intrinsic randomness, lead to data that may be better approximated as a stochastic process. However, these systems may require more sophisticated tools of analysis [10]. In order to identify an SDE, we need to estimate a diffusion function, which determines the nature of the random forcing, in addition to the drift, which represents the mean dynamics. In the context of single particle tracking, the diffusion constant is locally estimated using the mean square displacement [26], and the uncertainty can be quantified and compared against the Cramér–Rao bound [24, 34].

The mean square displacement can be generalized for the estimation of SDEs, where the drift and diffusion functions may vary spatially. Local approximations for the drift and diffusion functions can be obtained from data using the Kramers–Moyal expansion, and can be used to estimate spatially varying parameters [31, 11, 7, 30]. An estimate of the diffusion parameter which improves upon the Kramers–Moyal estimate is given in [27]. Further improvements have been made, such as estimates of the diffusion that are unbiased in the presence of measurement noise [34, 12] and methods which allow for the estimation of underdamped Langevin equations [2].

The estimation of the drift and diffusion functions using these Kramers–Moyal estimates extends naturally into the SINDy framework. Stochastic force inference, as presented in [12], is a similar nonparametric identification method for SDEs, which differs in that it does not use a sparse solver. In [1], the SINDy algorithm was used to estimate the parameters for an SDE using these Kramers–Moyal estimates. This method was expanded in [8]: solution methods based on binning and cross validation were introduced to reduce the effects of noise. Callaham et al. [5] expanded upon this method by adapting it to applications for which the random forcing cannot be considered white noise.

In the paper, we conduct a numerical analysis for using SINDy for stochastic systems and introduce improved methods which give higher order convergence. As previously mentioned, in [1] the drift and diffusion are approximated using the Kramers–Moyal formulas. We demonstrate the convergence rates of the algorithm with respect to the sampling period and the length of the trajectory. The approximations given in [1] only give first order convergence with respect to the sampling frequency. A similar analysis of the Kramers–Moyal estimates based on binning can be found in [6]. Additionally, since they only converge in expectation, we may require a long trajectory for the variance of the estimates to be tolerable. Combined, the high sampling frequency and long trajectories can make the data requirements to use SINDy for an SDE very demanding. To help remedy this, we demonstrate how we can develop higher order approximations of the drift and diffusion functions for use in SINDy.

The paper is organized as follows: First, we will review the SINDy algorithm and some concepts from SDEs which we will be using in this paper. We will then conduct a numerical analysis of the algorithms presented in [1], using the Ito–Taylor expansion of the SDE. Next,

we will present new, higher order methods and show the convergence rates of these methods. Finally, we will test all of these methods on several numerical examples to demonstrate how the new methods allow us to compute far more accurate approximations of the system for a given sampling frequency and trajectory length.

2. Sparse identification of nonlinear dynamics (SINDy).

2.1. Overview.

Consider a system governed by the ODE

$$\dot{x} = f(x), \qquad x \in \mathbb{R}^d. \tag{2.1}$$

If the dynamics of the system, f, are unknown, we would like to be able to estimate the function f using only data from the system. The SINDy algorithm [3] estimates f by choosing a dictionary of functions, $\theta = [\theta_1, \theta_2, \dots, \theta_k]$, and assuming f can be expressed (or approximated) as a linear combination of these functions. The ith component of f, f_i, can then be expressed as

$$f_i(x) = \sum_{j=1}^{k} \theta_j(x)\alpha_{i,j} = \theta(x)\alpha_i,$$

where $\theta = [\theta_1 \ \cdots \ \theta_k]$ is a row vector containing the dictionary functions and $\alpha^i = [\alpha_1^i \ \cdots \ \alpha_k^i]^T$ is the column vector of coefficients. Given data for $f(x_j)$ and $\theta(x_j)$ for $j = 1, \dots, n$, we can find the coefficients α_i by solving the minimization

$$\alpha_i = \underset{v}{\operatorname{argmin}} \sum_{j=1}^{n} |f_i(x_j) - \theta(x_j)v|^2. \tag{2.2}$$

This optimization can be solved by letting

$$\Theta = \begin{bmatrix} \theta(x_1) \\ \theta(x_2) \\ \vdots \\ \theta(x_n) \end{bmatrix}, \quad F = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad \text{and} \quad \alpha = \begin{bmatrix} \alpha^1 & \alpha^2 & \cdots & \alpha^d \end{bmatrix},$$

and computing $\alpha = \Theta^+ F$.

2.2. Approximating $f(x)$.

Typically, data for $f(x)$ cannot be measured directly. Instead, it is usually approximated using finite differences. The forward difference gives us a simple, first order approximation to f:

$$f(x(t)) = \frac{x(t + \Delta t) - x(t)}{\Delta t} + O(\Delta t). \tag{2.3}$$

Here $O(\Delta t)$ is the Landau "big O" notation. The approximation (2.3) is derived from the Taylor expansion of x,

$$x(t + \Delta t) = x(t) + \dot{x}(t)\Delta t + \ddot{x}(t)\frac{\Delta t^2}{2} + \cdots = x(t) + f(x(t))\Delta t + \frac{\partial f}{\partial x}\Big|_{x(t)} f(x(t))\frac{\Delta t^2}{2} + \cdots, \tag{2.4}$$

for f sufficiently smooth. The Taylor expansion (2.4) is also used to derive higher order methods, such as the central difference,

$$f(x) = \frac{x(t+\Delta t) - x(t-\Delta t)}{2\Delta t} + O(\Delta t^2). \tag{2.5}$$

We can use these finite differences to populate the matrix F used in the optimization (2.2), knowing that we can control the error with a small enough step size.

2.3. Sparse solutions. Since we choose an arbitrary dictionary of functions, $\{\theta_1, \dots, \theta_k\}$, the conditioning of the minimization (2.2) can become very poor. Additionally, if the dictionary is large and contains many redundant functions, having a solution which contains only a few nonzero entries would help to provide a simple interpretable result. The SINDy algorithm addresses these by using a sparse solution to (2.2). There are multiple methods for obtaining a sparse solution such as the least absolute shrinkage and selection operator (LASSO) or the sequentially thresholded least squares algorithm [3]. Using a sparse solution will give us a simpler identified system and improves the performance over the least squares solution.

3. Review of SDEs.

Consider the Ito SDE

$$dX_t = \mu(X_t)dt + \sigma(X_t)dW_t, \tag{3.1}$$

where $X_t \in \mathbb{R}^d$ and W_t is d-dimensional Brownian motion. The function $\mu : \mathbb{R}^d \to \mathbb{R}^d$ is the drift, a vector field which determines the average motion of system, while $\sigma : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is the diffusion function, which governs the stochastic forcing. The diffusion, σ, is also assumed to be positive definite. Motivated by SINDy, we wish to estimate μ and σ^2 from data. We note that we are estimating $\Sigma = \frac{1}{2}\sigma^2$ and not σ directly. However, if σ is positive definite, which is assumed, σ^2 uniquely determines σ.

3.1. Ergodicity. Since SINDy represents functions using the data vectors evaluated along the trajectory, we will need to relate the data vectors to the functions represented in some function space. To do this, we will assume that the process X_t has an ergodic measure ρ, so that both

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_t)dt = \int_{\mathbb{R}^d} f(x)d\rho(x) \qquad \text{and} \qquad \lim_{N\to\infty} \frac{1}{N}\sum_{i=0}^{N-1} f(X_{t_i}) = \int_{\mathbb{R}^d} f(x)d\rho(x) \tag{3.2}$$

hold almost surely. Some sufficient conditions that ensure that the SDE (3.1) generates a process with a stationary or an ergodic measure are given in, e.g., [16].

With this ergodic measure, the natural function space to consider is the Hilbert space $L^2(\rho)$. For any two functions $f, g \in L^2(\rho)$, we can use time averages to evaluate inner products:

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T g^*(X_t)f(X_t)dt = \lim_{N\to\infty} \frac{1}{N}\sum_{i=0}^{N-1} g^*(X_{t_n})f(X_{t_n}) = \int_{\mathbb{R}^d} g^* f\, d\rho = \langle f, g\rangle. \tag{3.3}$$

For notational simplicity, we will also use the brackets $\langle\cdot,\cdot\rangle$ to denote the matrix of inner products for two row vector-valued functions: if $f = \begin{bmatrix} f_1 & \cdots & f_k \end{bmatrix}$ and $g = \begin{bmatrix} g_1 & \cdots & g_l \end{bmatrix}$,

$$\langle f, g\rangle^{i,j} = \langle f^j, g^i\rangle, \qquad \text{or equivalently,} \qquad \langle f, g\rangle = \int_{\mathbb{R}^d} g^* f\, d\rho.$$

3.2. Ito–Taylor expansion. In order to evaluate the performance of different SINDy methods on SDEs, we will need to use the Ito–Taylor expansion of the solution. Let $\Sigma = \frac{1}{2}\sigma^2$. Following the notation of [17], let

$$L^0 = \sum_{j=1}^{d} \mu^j \frac{\partial}{\partial x^j} + \sum_{j,l}^{d} (\Sigma)^{j,l} \frac{\partial^2}{\partial x^j \partial x^l}$$

be the operator for the Ito equation (3.1) and define the operators

$$L^j = \sum_{i=1}^{d} \sigma^{i,j} \frac{\partial}{\partial x^i}.$$

These operators will give us the coefficients for the Ito–Taylor expansion of a function f. Denoting $\Delta W_t^i = W_{t+\Delta t}^i - W_t^i$, the first couple of terms are

$$f(X_{t+\Delta t}) = f(X_t) + L^0 f(X_t)\Delta t + \sum_{i=1}^{d} L^i f(X_t)\Delta W_t^i + (L^0)^2 f(X_t)\Delta t + \sum_{i=1}^{d} L^i L^0 f(X_t) \int_t^{t+\Delta t} \int_t^{s_1} dW_{s_2}^i ds_1 + \sum_{i=1}^{d} L^0 L^i f(X_t) \int_t^{t+\Delta t} \int_t^{s_1} ds_2 dW_{s_1}^i + \cdots.$$

The general Ito–Taylor expansions can be found in Theorem 5.5.1 of [17]. We will use the Ito–Taylor expansion to develop estimates for μ^i and $\sigma^{i,j}$. For the purposes of this paper, we will be able to specialize to a few cases, which will allow us to quantify the error in our estimates while also being simpler to manipulate than the larger expansion.

3.2.1. Weak expansion. The first specialization of the Ito–Taylor expansion will be a weak expansion, which will allow us to estimate the expected error in our estimate:

$$\mathbb{E}(f(X_{t+\Delta t})|X_t) = f(X_t) + \sum_{m=1}^{k} (L^0)^m f(X_t) \frac{\Delta t^m}{m!} + R(X_t) \tag{3.4}$$

with $R(X_t) = O(\Delta t^{m+1})$.

This expansion follows from Proposition 5.5.1 and Lemma 5.7.1 of [17]. Theorem 5.5.1 gives the general Ito–Taylor expansion, while Lemma 5.7.1 shows that all multiple Ito integrals which contain integration with respect to a component of the Wiener process have zero first moment. The remainder term is then a standard integral.

We will consider the expansion (3.4) with the functions $f(x) = x^i$ to get

$$\mathbb{E}(X_{t+\Delta t}^i|X_t) = X_t^i + \mu^i(X_t)\Delta t + \sum_{m=2}^{k} (L^0)^{m-1} \mu^i(X_t) \frac{\Delta t^m}{m!} + O(\Delta t^{k+1}) \tag{3.5}$$

to estimate the drift. To estimate the diffusion, we will let $f(x) = (x^i - X_t^i)(x^j - X_t^j)$, with X_t held constant at the value at the beginning of the time step, to get

$$\mathbb{E}(f(X_{t+\Delta t}) \,|\, X_t) = 2\Sigma^{i,j}(X_t)\Delta t + g(X_t)\Delta t^2 + O(\Delta t^3), \tag{3.6}$$

where

$$g = L^0\Sigma^{i,j} + \mu^i\mu^j + \sum_{k=1}^{d}\left(\Sigma^{i,k}\frac{\partial \mu^j}{\partial x^k} + \Sigma^{j,k}\frac{\partial \mu^i}{\partial x^k}\right).$$

3.2.2. Strong expansions. We will also use the strong Ito–Taylor expansion, which will give a bound on the variance of our estimates. These immediately follow from Proposition 5.9.1 of [17]. First, if we apply it to $f(x) = x^i$, we have

$$X^i_{t+\Delta t} - X^i_t = \mu^i(X_t)\Delta t + \sum_{m=1}^{d}\sigma^{i,m}(X_t)\Delta W^m_t + R_t, \tag{3.7}$$

where $\mathbb{E}(|R_t|^2|X_t)d\rho = O(\Delta t^2)$.

Similarly, we can apply the same proposition to $f(x) = (x^i - X^i_t)(x^j - X^j_t)$, which gives us (after moving around some of the terms)

$$(X^i_{t+\Delta t} - X^i_t)(X^j_{t+\Delta t} - X^j_t) = 2\Sigma^{i,j}(X_t)\Delta t + \sum_{k,l=1}^{d}(\sigma^{k,i}\sigma^{l,j}(X_t) + \sigma^{k,j}\sigma^{l,i}(X_t))I_{(i,j)} + R_t, \tag{3.8}$$

where $\mathbb{E}(|R_t|^2|X_t) = O(\Delta t^3)$ and $I_{(i,j)} = \int_0^{\Delta t}\int_0^{s_1} dW^i_{s_2}dW^j_{s_1}$. When we create estimates of $\mu^i(X_t)$ and $\Sigma^{i,j}(X_t)$, the expansions (3.7) and (3.8) will be useful in bounding the variance of these two estimates.

Remark 1. For the expansions, it is implicit that we must assume that all (up to the necessary order) of the coefficient functions, $L^{a_1}L^{a_2}\cdots L^{a_n}f$, satisfy the requirements with respect to the multiple Ito integrals set forth in Chapter 5 of [17]. The conditions set forth are necessary for the Ito–Taylor expansions to be valid locally.

Additionally, we will also need to assume that the remainder terms will be square integrable with respect to the ergodic measure. In particular, we will assume

$$\int_{\mathbb{R}^d}|R(x)|^2 d\rho(x) = O(\Delta t^{m+1})$$

in the weak expansion and

$$\int_{\mathbb{R}^d} R_2(x)^2 d\rho(x) = O(\Delta t) \quad \big(\text{or} \quad O(\Delta t^2)\big)$$

in the strong expansions, where $R_2(x) = \mathbb{E}(|R_t|^2 \mid X_t = x)$. This assumption will allow us to take time averages and expect them to be finite. Following the proofs in [17], it can be seen that these can be guaranteed by imposing similar integrability conditions on the coefficient functions with respect to the ergodic measure. This will often be the case, as the ergodic measure will decay rapidly toward infinity. A sufficiently strong condition to guarantee the integrability of the error is, for example, that both the diffusion and drift functions are smooth and the derivatives of all orders are bounded.

4. SINDy for stochastic systems. Given data for the drift and diffusion matrix of (3.1), we can set up an optimization problem similar to (2.2). Similar to the deterministic case, we can also approximate μ and Σ using finite differences. As before, we assume that we have a dictionary $\theta = [\theta_1, \theta_2, \dots, \theta_k]$ and that each of the components of μ and Σ lies in the span of the components of θ:

$$\mu^i = \theta\alpha^i \qquad \text{and} \qquad \Sigma^{i,j} = \theta\beta^{i,j}.$$

Suppose we have the data from a trajectory of length T with sampling period Δt. If we let $\Delta X^i_{t_n} = X^i_{t_{n+1}} - X^i_{t_n}$, we can approximate the drift using

$$\mu^i(X_{t_m}) \approx \frac{X^i_{t_{m+1}} - X^i_{t_m}}{\Delta t} = \frac{\Delta X^i_t}{\Delta t}. \tag{4.1}$$

Similarly, we can approximate the diffusion with

$$\Sigma^{i,j}(X_{t_m}) \approx \frac{(X^i_{t_{m+1}} - X^i_{t_m})(X^j_{t_{m+1}} - X^j_{t_m})}{2\Delta t} = \frac{\Delta X^i_{t_m}\Delta X^j_{t_m}}{2\Delta t}. \tag{4.2}$$

It was shown in [1] that we can use the approximations (4.1) and (4.2) to set up the minimization problems

$$\tilde\alpha^i = \operatorname*{argmin}_v \sum_{m=0}^{N-1} \left| \frac{\Delta X^i_{t_m}}{\Delta t} - \theta(X_{t_m})v \right|^2 \tag{4.3}$$

and

$$\tilde\beta^{i,j} = \operatorname*{argmin}_v \sum_{m=0}^{N-1} \left| \frac{\Delta X^i_{t_m}\Delta X^j_{t_m}}{2\Delta t} - \theta(X_{t_m})v \right|^2. \tag{4.4}$$

Under the assumptions set forth in Remark 1, we can show that as $\Delta t \to 0$ and $T \to \infty$, the coefficients given by (4.3) and (4.4) converge to the true coefficients: $\tilde\alpha^i \to \alpha^i$ and $\tilde\beta^{i,j} \to \beta^{i,j}$.

If we define the matrices

$$\Theta = \begin{bmatrix} \theta(X_{t_0}) \\ \theta(X_{t_1}) \\ \vdots \\ \theta(X_{t_{N-1}}) \end{bmatrix} \qquad \text{and} \qquad D^i = \begin{bmatrix} \Delta X^i_{t_0} \\ \Delta X^i_{t_1} \\ \vdots \\ \Delta X^i_{t_{N-1}} \end{bmatrix}, \tag{4.5}$$

we can express (4.3) and (4.4) concisely as

$$\tilde\alpha^i = \operatorname*{argmin}_v \left\| \frac{D^i}{\Delta t} - \Theta v \right\| \qquad \text{and} \qquad \beta^{i,j} = \operatorname*{argmin}_v \left\| \frac{D^i \odot D^j}{2\Delta t} - \Theta v \right\|.$$

(Here $D^i \odot D^j$ represents the Hadamard, or elementwise, product.) These equations are solved by $\tilde\alpha_i = \Delta t^{-1}\Theta^+ D^i$ and $\tilde\beta_{i,j} = (2\Delta t)^{-1}\Theta^+(D^i \odot D^j)$, respectively.

Theorem 4.1. *Let X_t be an ergodic drift-diffusion process generated by the SDE (3.1). Consider the optimization problems (4.3) and (4.4) using data from a trajectory of length T sampled with frequency Δt. Suppose that the components of θ are linearly independent and span the subspace $\mathcal{F}$, and that the assumptions on the Ito–Taylor expansions outlined in Remark 1 are met. If μ^i or $\Sigma^{i,j}$ lie in $\mathcal{F}$, then the vectors given by corresponding optimization converge in probability to the true coefficients as $T \to \infty$ and $\Delta t \to 0$. That is, $\tilde{\alpha}^i \to \alpha^i$ or $\tilde{\beta}^{i,j} \to \beta^{i,j}$.*

Theorem 4.1 was shown in [1] and will be implied by the stronger Theorems 5.1 and 5.2 which give rates of convergence. However, we will demonstrate the main reasoning behind the proof, as it will be informative to our later analysis. By the assumptions of Theorem 4.1 we have that Θ has full rank and $\mu = \theta\alpha^i$, $\Sigma^{i,j} = \theta\beta^{i,j}$.

$$\tilde{\alpha}^i = (\Theta^*\Theta)^{-1}\Theta^*\frac{D^i}{\Delta t} = \left(\frac{1}{N}\Theta^*\Theta\right)^{-1}\left(\frac{1}{N\Delta t}\Theta^* D^i\right),$$

where $N = T/\Delta t$ is the number of data samples. The first quantity can be evaluated using ergodicity as $N \to \infty$:

$$\frac{1}{N}\Theta^*\Theta = \frac{1}{N}\sum_{m=0}^{N-1}\theta^*(X_{t_m})\theta(X_{t_m}) \xrightarrow{N} \langle\theta,\theta\rangle.$$

For the second expression, the definition of the stochastic integral gives us

$$\Theta^* D^i = \sum_{m=0}^{N-1}\theta^*(X_m)(X^i_{t_{m+1}} - X^i_{t_m}) \xrightarrow{\Delta t} \int_{t_0}^{t_0+T}\theta^* dX^i$$

as $\Delta t \to 0$. Finally, using (3.1) and (3.3), we can show

$$\frac{1}{N\Delta t}\Theta^* D^i \xrightarrow{\Delta t} \frac{1}{T}\int_{t_0}^{t_0+T}\theta^* dX^i \xrightarrow{T} \langle\mu,\theta\rangle = \langle\theta,\theta\rangle\alpha^i \tag{4.6}$$

as $\Delta t \to 0$ and $T \to \infty$. The limit as $\Delta t \to 0$ gives the convergence of the sum to the stochastic integral, and the limit as $T \to \infty$ allows us to sample almost everywhere on the stationary measure for the ergodic convergence. Similarly, we can use the convergence

$$\sum_{m=0}^{N-1}\theta^*(X_{t_m})(X^i_{t_{m+1}} - X^i_{t_m})(X^j_{t_{m+1}} - X^j_{t_m}) \xrightarrow{\Delta t} \int_{t_0}^{t_0+T}\theta^* d[X^i, X^j], \qquad \Delta t \to 0$$

to show that $\frac{1}{2N\Delta t}\Theta^*(D^i \odot D^j) \to \langle\Sigma^{i,j},\theta\rangle = \langle\theta,\theta\rangle\beta^{i,j}$. (Here $[X,Y]_t$ is the quadratic covariation process of X_t, and Y_t.) This would establish the result, except that we used the iterated limits $\Delta t \to 0$ and $T \to \infty$ in (4.6) without showing the double limit exists. This is where we would use the integrability assumptions in Remark 1, which are used in the proofs of Theorems 5.1 and 5.2.

Theorem 4.1 demonstrates how the least squares solutions converge to the true coefficients of the SDE. However, the SINDy algorithm finds a sparse solution, which can greatly improve

the accuracy of the results over the least squares solution. To set this up, the two optimizations (4.3) and (4.4) can be summarized using the normal equations,

$$\Theta^*\Theta\tilde{\alpha}^i = \frac{1}{\Delta t}\Theta^* D^i \tag{4.7}$$

and

$$\Theta^*\Theta\tilde{\beta}^{i,j} = \frac{1}{2\Delta t}\Theta^*(D^i \odot D^j). \tag{4.8}$$

We can then solve (4.7) and (4.8) using a sparse solver, such as the one proposed in [3] to obtain a sparse solution.

5. Numerical analysis of stochastic SINDy. Theorem 4.1 claims that as $\Delta t \to 0$ and $T \to \infty$, the coefficients given by (4.3) and (4.4) converge to the true parameters of the SDE (3.1) as $\Delta t \to 0$ and $T \to \infty$. However, for real experiments, there will be limits to the sampling frequency and the length of trajectory for which we can acquire data. In [12], the trajectory of the SDE was interpreted as a noisy transmission channel, and estimates on the relative squared errors were derived based on the information content of the signal.

In this section, we will use an alternate approach of deriving the error in the estimate based on the Ito–Taylor expansion of the SDE. We will look at both the bias and variance of the approximations for finite Δt and T. In particular, we derive the error with explicit constants (up to the leading order) in terms of the dictionary θ and the functions μ and σ.

In this setting, we will be using both "big O" and "little o" notation. The "big O" notation will be used to denote convergence as $\Delta t \to 0$. These terms will come from the higher order error terms in the estimators of $\mu^i(X_t)$ and $\Sigma^{i,j}(X_t)$. In particular, the constant in the "big O" will depend only on the parameters of the SDE; it does not depend on the initial condition, trajectory length, or realization of the trajectory.

The "little o" will denote convergence with respect to T. Specifically $o(1)$ denotes a function that goes to zero as $T \to \infty$. This will capture the ergodic convergence; the $o(1)$ term will be the error that comes from the finite trajectory failing to completely sample the ergodic measure.

The SINDy algorithm will give us vectors of coefficients, $\tilde{\alpha}^i$ and $\tilde{\beta}^{i,j}$, for the system. We will be interested in the error of these vectors relative to the true coefficients α^i and $\beta^{i,j}$,

$$err = \tilde{\alpha}^i - \alpha^i \quad \text{or} \quad err = \tilde{\beta}^{i,j} - \beta^{i,j}.$$

(We note that this error is specifically for the vector α^i or $\beta^{i,j}$ being estimated, even though it is not indexed. Since each vector is estimated separately, there should be no confusion.) This error will be a random variable depending on the realization of the system. To evaluate the performance of the algorithms, we will use the mean and variance of this error:

$$err_{mean} = \|\mathbb{E}(err)\|_2 \quad \text{and} \quad err_{var} = Var(err) = \mathbb{E}(\|err - \mathbb{E}(err)\|_2^2).$$

The mean and variance of the error measure the bias and spread in the estimates $\tilde{\alpha}^i$ and $\tilde{\beta}^{i,j}$. These errors in the coefficients can be quantified using the errors in the estimates of μ^i and $\Sigma^{i,j}$ given in (4.1) and (4.2) at each step. We will present the analysis for the drift coefficients, α^i, noting that analysis for the diffusion follows the same path.

5.1. Drift. As mentioned, the error in $\tilde{\alpha}^i$ stems from the error in the approximation in (4.1)

$$\mu^i(X_{t_n}) \approx \frac{X_{t_{n+1}} - X_{t_n}}{\Delta t}.$$

We can define the error

$$e_{t_n} = \frac{X^i_{t_{n+1}} - X^i_{t_n}}{\Delta t} - \mu^i(X_{t_n}).$$

The order of the error, e_t, at each time step will directly determine the error in the coefficients $\tilde{\alpha}^i$. We can use Ito–Taylor expansions for X_t to bound both $\mathbb{E}(|e_t|)$ and $\mathbb{E}(|e_t|^2)$. The weak Ito–Taylor expansion (3.4) gives us

(5.1)
$$\mathbb{E}(e_t \,|\, X_t) = \frac{1}{\Delta t}\left(\mu^i(X_t)\Delta t + L^0\mu^i(X_t)\frac{\Delta t^2}{2} + O(\Delta t^3)\right) - \mu^i(X_t) = L^0\mu^i(X_t)\frac{\Delta t}{2} + O(\Delta t^2).$$

Similarly, we can use the strong truncation (3.7) to obtain

$$e_t = \sum_{m=1}^{d} \sigma^{i,m}(X_t)\frac{\Delta W^m_t}{\Delta t} + \frac{R_t}{\Delta t},$$

where $\mathbb{E}(|R_t|^2|X_t) = O(\Delta t^2)$. Then, taking the expectance of e_t^2, we get

$$\mathbb{E}(|e_t|^2 \,|\, X_t) = \sum_{m=1}^{d} \frac{\sigma^{i,m}(X_t)^2}{\Delta t} + O\left(\Delta t^{\frac{-1}{2}}\right). \tag{5.2}$$

Now, let E be the matrix containing the time samples of e_t,

$$E = \begin{bmatrix} e_{t_0} & e_{t_1} & \cdots & e_{t_{N-1}} \end{bmatrix}^T = \frac{D^i}{\Delta t} - \Theta\alpha^i,$$

using $\theta(X_t)\alpha^i = \mu^i(X_t)$. Then we have

$$err = \tilde{\alpha}^i - \alpha^i = \Theta^+\frac{D^i}{\Delta t} - \Theta^+\Theta\alpha = (\Theta^*\Theta)^{-1}\Theta^* E. \tag{5.3}$$

Using ergodicity, we have

$$\left(\frac{1}{N}\Theta^*\Theta\right)^{-1} = (\langle\theta,\theta\rangle + o(1))^{-1} = \langle\theta,\theta\rangle^{-1} + o(1), \tag{5.4}$$

which allows us to evaluate the first term in (5.3):

$$err = (\langle\theta,\theta\rangle^{-1} + o(1))\left(\frac{1}{N}\Theta^* E\right). \tag{5.5}$$

Bounding the mean and variance will follow from bounds on the mean and variance of $\frac{1}{N}\Theta^* E$.

Theorem 5.1. *Consider the optimization problem given by (4.1) and (4.3). Then the bias is bounded by*

$$err_{mean} \leq \frac{C_1}{2}\left(\|L^0\mu^i\|_2 + O(\Delta t) + o(1)\right)\Delta t$$

and

$$err_{var} \leq \frac{C_2}{T}\left(\sum_{m=1}^{d}\|\sigma^{i,m}\|_4^2 + O\left(\Delta t^{\frac{1}{2}}\right) + o(1)\right),$$

where

$$C_1 = \|\langle\theta,\theta\rangle^{-1}\|_2\|\theta\|_2 \quad \text{and} \quad C_2 = \|\langle\theta,\theta\rangle^{-1}\|_2^2\|\theta\|_4^2 \tag{5.6}$$

depend only on the choice of θ.

As stated in Theorem 5.1, in expectation, the accuracy of our estimate depends primarily on the sampling period Δt, and not on the length of the trajectory. The length of the trajectory instead controls the variance of the estimate, which is proportional to $1/T$. Up to the leading term, the variance does not depend on the sampling period. These results previously appeared in [12, 2], although our proof is different. This pattern will persist as we develop higher order methods for estimating the drift, where the sampling frequency determines the bias and the length of the trajectory determines the variance.

Proof. For the mean error, we will need to bound the quantity $\frac{1}{N}\|\mathbb{E}(\Theta^* E)\|$. We have

$$\mathbb{E}\left(\frac{1}{N}\Theta^* E\right) = \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1}\theta^*(X_{t_n})e_{t_n}\right) = \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1}\theta^*(X_{t_n})\mathbb{E}(e_{t_n}\,|\,X_{t_n})\right).$$

Then, using ergodicity and (5.1), we obtain

$$\begin{aligned}\mathbb{E}\left(\frac{1}{N}\Theta^* E\right) &= \mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1}\theta^*(X_{t_n})\left(\frac{\Delta t}{2}L^0\mu^i(X_{t_n}) + O(\Delta t^2)\right)\right)\\ &= \frac{\Delta t}{2}\left(\langle L^0\mu^i,\theta\rangle + o(1)\right) + O(\Delta t^2).\end{aligned}$$

Finally, using (5.5), we get

$$\begin{aligned}\|\mathbb{E}(err)\| &= \left\|\left(\langle\theta,\theta\rangle^{-1} + o(1)\right)\right\|_2\left(\frac{\Delta t}{2}\left(\langle L^0\mu^i,\theta\rangle + o(1)\right) + O(\Delta t^2)\right)\\ &\leq \|\langle\theta,\theta\rangle^{-1}\|_2\left(\|\theta\|_2\|L^0\mu^i\|_2 + O(\Delta t) + o(1)\right)\frac{\Delta t}{2} = C_1\left(\|L^0\mu^i\|_2 + O(\Delta t) + o(1)\right)\frac{\Delta t}{2}.\end{aligned}$$

This bounds the mean error. To find the variance, we have

$$\begin{aligned}Var\left(\frac{1}{N}\Theta^* E\right) &\leq \mathbb{E}\left(\left\|\frac{1}{N}\Theta^* E\right\|_2^2\right) = \mathbb{E}\left(\left\|\sum_{n=0}^{N-1}\theta^*(X_{t_n})e_{t_n}\right\|_2^2\right) \leq \mathbb{E}\left(\sum_{n=0}^{N-1}\|\theta^*(X_{t_n})\|_2^2|e_{t_n}|^2\|\right)\\ &= \mathbb{E}\left(\sum_{n=0}^{N_1}\|\theta(X_{t_n})\|_2^2\,\mathbb{E}\left(|e_{t_n}|^2\,|\,X_{t_n}\right)\right).\end{aligned}$$

Now, using (5.2) with this equation, we have

$$\begin{aligned} Var\left(\frac{1}{N}\Theta^* E\right) &\le \mathbb{E}\left(\frac{1}{N^2}\sum_{n=0}^{N-1}\|\theta(X_{t_n})\|_2^2\left(\sum_{m=1}^{d}\frac{|\sigma^{i,m}|^2}{\Delta t}+O\left(\Delta t^{\frac{-1}{2}}\right)\right)\right) \\ &= \frac{1}{N\Delta t}\left(\sum_{m=1}^{d}\langle(\sigma^{i,m})^2,\|\theta\|_2^2\rangle+O\left(\Delta t^{\frac{1}{2}}\right)+o(1)\right) \\ &\le \frac{1}{T}\|\theta\|_4^2\left(\sum_{m=1}^{d}\|\sigma^{i,m}\|_4^2+O\left(\Delta t^{\frac{1}{2}}\right)+o(1)\right). \end{aligned}$$

Then

$$\begin{aligned} Var(err) &= \left(\|\langle\theta,\theta\rangle^{-1}\|_2^2+o(1)\right)\|\theta\|_4^2\left(\frac{1}{T}\left(\sum_{m=1}^{d}\|\sigma^{i,m}\|_4^2+O\left(\Delta t^{\frac{1}{2}}\right)+o(1)\right)\right) \\ &= \frac{\|\langle\theta,\theta\rangle^{-1}\|_2^2\|\theta\|_4^2}{T}\left(\sum_{m=1}^{d}\|\sigma^{i,m}\|_4^2+O\left(\Delta t^{\frac{1}{2}}\right)+o(1)\right) \\ &= \frac{C_2}{T}\left(\sum_{m=1}^{d}\|\sigma^{i,m}\|_4^2+O\left(\Delta t^{\frac{1}{2}}\right)+o(1)\right). \end{aligned}$$

■

5.2. Diffusion. The analysis of the diffusion coefficients follows the same argument. The approximation for $\Sigma^{i,j}$ given in (4.2) is

$$\Sigma^{i,j}(X_{t_m})\approx\frac{(X^i_{t_{m+1}}-X^i_{t_m})(X^j_{t_{m+1}}-X^j_{t_m})}{2\Delta t}=\frac{\Delta X^i_{t_m}\Delta X^j_{t_m}}{2\Delta t}.$$

Then we can define the error

$$e_t=\frac{(X^i_{t+\Delta t}-X^i_t)(X^j_{t+\Delta t}-X^j_t)}{2\Delta t}-\Sigma^{i,j}(X_t).$$

We can use the weak Ito–Taylor expansion (3.6) to bound $\mathbb{E}(e_t\,|\,X_t)$:

$$\mathbb{E}(e_t\,|\,X_t)=g(X_t)\frac{\Delta t}{2}+O(\Delta t^2),\qquad g=L^0\Sigma^{i,j}+\mu^i\mu^j+\sum_{k=1}^{d}\left(\Sigma^{i,k}\frac{\partial\mu^j}{\partial x^k}+\Sigma^{j,k}\frac{\partial\mu^i}{\partial x^k}\right). \tag{5.7}$$

Similarly, the strong Ito–Taylor expansion (3.8) gives us (see Appendix A.2)

$$\mathbb{E}(|e_t|^2\,|\,X_t)=\Sigma^{i,i}(X_t)\Sigma^{j,j}(X_t)+\Sigma^{i,j}(X_t)^2+O(\Delta t^{\frac{1}{2}}). \tag{5.8}$$

Theorem 5.2. *Consider the optimization problem given by* (4.2) *and* (4.4). *Then the mean error is bounded by*

$$err_{mean}=\frac{C_1}{2}(\|g\|+O(\Delta t)+o(1))\Delta t,$$

where

$$g = L^0\Sigma^{i,j} + \mu^i\mu^j + \sum_{k=1}^{d}\left(\Sigma^{i,k}\frac{\partial\mu^j}{\partial x^k} + \Sigma^{j,k}\frac{\partial\mu^i}{\partial x^k}\right).$$

The variance is bounded by

$$err_{var} = \frac{C_2}{4}\left(\left\|\Sigma^{i,i}\Sigma^{j,j} + (\Sigma^{i,j})^2\right\| + O(\Delta t^{\frac{1}{2}}) + o(1)\right)\frac{\Delta t}{T}.$$

The constants C_1 and C_2 are the same as those given in (5.6).

Proof. The proof follows that of Theorem 5.1, except using (5.7) and (5.8) to bound $|\mathbb{E}(e_t \,|\, X_t)|$ and $\mathbb{E}\left(|e_t|^2 \,|\, X_t\right)$, respectively. ∎

Similar to Theorem 5.1, the argument above shows that the mean error converges with order Δt. However, unlike the estimate for the drift, when estimating the diffusion the variance is proportional to both Δt and $1/T$. Similar to the drift, these results were shown previously in [12, 2]. Later, we will see that the higher order estimates for the diffusion will also have variance proportional to $\Delta t/T$.

6. Higher order methods. From Theorems 5.1 and 5.2 we can see that the quantities Δt, T, C_1, and C_2 will control the magnitude of the error. The constants, C_1 and C_2, depend only on the choice of the dictionary θ, which determines the conditioning of the problem. The SINDy algorithm also uses a sparsity promoting algorithm which can improve the conditioning of the problem and force many of the coefficients to zero, which can reduce the error [3, 1]. However, even if the sparsity promoting algorithm chooses all of the correct coefficients, we have just shown that there is still a limit to the accuracy of the estimation determined by the sampling frequency and trajectory. The primary purpose of this section is to analyze alternate methods of approximating μ^i and $\Sigma^{i,j}$ which can improve the performance of SINDy (with respect to Δt).

The methods above resulted from first order approximations (4.1) and (4.2) of $\mu^i(X_t)$ and $\Sigma^{i,j}(X_t)$, respectively. Higher order approximations of these data points can in turn lead to more accurate approximations of the functions in the output of SINDy. We can generate better approximations for the drift using multistep difference methods. The use of linear multistep methods (LMMs) to estimate dynamics is investigated in [15] for deterministic systems. While the estimates for the diffusion will be similar, they cannot be achieved strictly using LMMs.

In order to achieve a higher order approximation, we will need to use more data points in the approximation at each time step. As such, we will define

$$\Theta_n = \begin{bmatrix} \theta(X_{t_n}) \\ \theta(X_{t_{n+1}}) \\ \vdots \\ \theta(X_{t_{N+n-1}}) \end{bmatrix} \quad \text{and} \quad D_n^i = \begin{bmatrix} X^i_{t_n} - X^i_{t_0} \\ X^i_{t_{n+1}} - X^i_{t_1} \\ \vdots \\ X^i_{t_{N+n-1}} - X^i_{t_{N-1}} \end{bmatrix}. \tag{6.1}$$

With this definition, Θ_n contains the data of θ time delayed by n steps. With the earlier definition of Θ, we have $\Theta = \Theta_0$. Similarly, D_n^i contains the data for the change in X over n time steps, with $D_1^i = D^i$ using the earlier definition of D^i.

6.1. Drift. First, we will look to make improvements on estimating the drift. These estimates will be simpler than those for the diffusion. As mentioned, these approximations are directly analogous to the LMMs used in the simulation of deterministic systems.

6.1.1. Second order forward difference. The first order forward difference, which is used to approximate μ^i in Theorem 5.1, is also commonly used to approximate the derivative $f(x)$ in the differential equation $\dot{x} = f(x)$. In fact, if we compare the weak Ito–Taylor expansion (3.4) with the deterministic Taylor series for an ODE, (2.4), we see that they are almost identical. There are many higher order methods which are used to approximate f in the simulation of ODEs. By analogy, we can expect that these methods would give an approximation of the same order for μ^i (in expectation). One of the simplest of these would be the second order forward difference,

$$\mu^i(X_{t_n}) \approx \frac{4(X_{t_{n+1}} - X_t) - (X_{t_{n+2}} - X_t)}{2\Delta t} = \frac{-3X^i_{t_n} + 4X^i_{t_{n+1}} - X^i_{t_{n+2}}}{2\Delta t}. \tag{6.2}$$

Similarly to before we can define the error in this approximation to be

$$e_t = \frac{-3X^i_t + 4X^i_{t+\Delta t} - X^i_{t+2\Delta t}}{2\Delta t} - \mu^i(X_t).$$

Using the weak Ito–Taylor expansion (3.4), it is easy to see that

$$\mathbb{E}(e_{t_n} \,|\, X_{t_n}) = -\frac{(L^0)^2\mu^i(X_{t_n})}{3}\Delta t^2 + O(\Delta t^3), \tag{6.3}$$

which shows that this method does indeed give a second order approximation of μ. Using this approximation, we can set up a matrix formulation of (6.2):

$$\Theta_0\alpha^i \approx \frac{1}{2\Delta t}\left(4D^i_1 - D^i_2\right).$$

If we set up the normal equations, this becomes

$$\Theta_0^*\Theta_0\tilde{\alpha}^i = \frac{1}{2\Delta t}\Theta_0^*\left(4D^i_1 - D^i_2\right). \tag{6.4}$$

Theorem 6.1. *Consider the approximation $\tilde{\alpha}^i$ obtained from (6.4). The mean error is bounded by*

$$\|\mathbb{E}(err)\|_2 = \frac{C_1}{3}(\|(L^0)^2\mu^i\| + O(\Delta t) + o(1))\Delta t^2$$

and the mean squared error by

$$\mathbb{E}\left(\|(err)\|_2^2\right) = \frac{C_2}{T}\left(\sum_j^d \|\sigma^{i,j}\|_4^2 + O(\Delta t^{\frac{1}{2}}) + o(1)\right).$$

The constants C_1 and C_2 are the same as those given in (5.6).

The proof of Theorem 6.1 is similar to that of Theorem 5.1, but requires some extra algebraic manipulation, so it is included in Appendix A.1.

Remark 2. These methods can easily be generalized to higher order methods using higher order finite differences, as will be done in section 6.1.3. However, the least squares solution only yields correct results for forward differences. Other finite difference methods can cause certain sums to converge to the wrong stochastic integral. For example, a central difference approximation for μ^i,

$$\mu_t^i \approx \frac{X_{t+\Delta t}^i - X_{t-\Delta t}^i}{2\Delta t},$$

gives us $\Theta_1 \alpha^i \approx \frac{1}{2\Delta t} D_2^i$. The normal equation for the least squares solution

$$\Theta_1^* \Theta_1 \tilde{\alpha}^i = \frac{1}{2\Delta t} \Theta_1^* D_2^i \tag{6.5}$$

gives the wrong results, because as $\Delta t \to 0$, $\frac{1}{2}\Theta_1^* D_2^i$ converges to the Stratonovich integral instead of the Ito integral,

$$\frac{1}{2}\Theta_1^* D_2^i \to \int_0^T \theta^*(X_t) \circ dX_t^i \neq \int_0^T \theta^*(X_t)\, dX_t^i,$$

and $\tilde{\alpha}^i$ will not converge to the correct value. To prevent this, (6.5) can instead be solved using

$$\Theta_0^* \Theta_1 \tilde{\alpha}^i = \frac{1}{2\Delta t} \Theta_0^* D_2^i,$$

which gives the proper convergence.This amounts to using Θ_0 as a set of instrumental variables (see [28]).

6.1.2. Trapezoidal Method. The second order method above uses additional measurements of X_t^i to provide a more accurate estimate of μ^i. Alternatively, we can use multiple measurements of μ^i to better approximate the difference $X_{t+\Delta t}^i - X_t^i$. Consider the first order forward difference given by (4.1).

$$\mu^i(X_{t_n}) \approx \frac{X_{t_{n+1}}^i - X_{t_n}^i}{\Delta t}.$$

Theorem 5.1 used this difference to give an order Δt approximation of μ^i. However, it turns out that $\frac{1}{2}(\mu^i(X_t) + \mu^i(X_{t+\Delta t}))$ gives a much better approximation of this difference:

$$\frac{1}{2}\left(\mu^i(X_{t_n}) + \mu^i(X_{t_{n+1}})\right) \approx \frac{X_{t_{n+1}}^i - X_{t_n}^i}{\Delta t}. \tag{6.6}$$

We will call this approximation the trapezoidal approximation, since this is exactly the trapezoidal method used in the numerical simulation of ODEs. If we consider the error in this equation,

$$e_t = \frac{X_{t_{n+1}}^i - X_{t_n}^i}{\Delta t} - \frac{1}{2}\left(\mu^i(X_{t_n}) + \mu^i(X_{t_{n+1}})\right),$$

we can use the weak Ito–Taylor approximations of X_t and $\mu^i(X_t)$ to show that

$$\mathbb{E}(e_t \mid X_t) = -(L^0)^2\mu^i(X_t)\frac{\Delta t^2}{12} + O(\Delta t^3). \tag{6.7}$$

This not only gives us a second order method, with respect to Δt, but the leading coefficient for the error is much smaller (by a factor of 1/8) than the second order forward difference.

To set up the matrix formulation of (6.6), we have

$$\frac{1}{2}(\Theta_0 + \Theta_1)\,\alpha^i \approx \frac{1}{\Delta t} D_1^i. \tag{6.8}$$

We can multiply (6.8) by Θ_0^* on each side to obtain

$$\frac{1}{2}\Theta_0^*(\Theta_0 + \Theta_1)\tilde{\alpha}^i = \frac{1}{\Delta t}\Theta_0^* D_1^i. \tag{6.9}$$

We can use this equation analogously to the normal equation; we will solve for $\tilde{\alpha}^i$ either directly using matrix inversion or by using a sparse solver.

Remark 3. We note that we cannot solve (6.8) using least squares,

$$\tilde{\alpha}^i \neq \frac{2}{\Delta t}(\Theta_0 + \Theta_1)^+ D_1^i.$$

Similarly to Remark 2, this leads to sums converging to the wrong stochastic integral. In [12], a similar method was used which leverages the convergence to the Stratonovich integral to generate an approximation which better handles noise. The authors' method corrects for the Ito versus Stratonovich differently from the one presented here and requires an accurate estimate of the divergence of the diffusion function.

Theorem 6.2. *Consider the estimation $\tilde{\alpha}^i$ given by solving (6.9). The mean error is bounded by*

$$err_{mean} \leq C_1 \frac{\Delta t^2}{12}(\|(L^0)^2\mu^i\|_2 + O(\Delta t) + o(1))$$

and

$$err_{var} \leq \frac{C_2}{T}\left(\sum_{j=1}^{d} \|\sigma^{i,j}\|_2^2 + O(\Delta t^{\frac{1}{2}}) + o(1)\right).$$

Proof. Letting E be the matrix containing the samples of e_t, we have

$$\frac{1}{\Delta t} D_1^i = \frac{1}{2}(\Theta_0 + \Theta_1)\alpha^i + E.$$

Using this in (6.9) gives us

$$\frac{1}{2}\Theta_0^*(\Theta_0 + \Theta_1)\tilde{\alpha^i} = \frac{1}{2}\Theta_0^*(\Theta_0 + \Theta_1)\alpha^i + \Theta_0^* E,$$

so the error is

$$err = \tilde{\alpha}^i - \alpha^i = \left(\frac{1}{2}\Theta_0^*(\Theta_0 + \Theta_1)\right)^{-1} \Theta_0^* E.$$

Since $\mathbb{E}(\theta(X_{t+\Delta t})|X_t) = \theta(X_t) + O(\Delta t)$, we can use ergodicity to evaluate

$$\frac{1}{2N}\Theta_0^*(\Theta_0 + \Theta_1) \to \langle \theta, \theta \rangle + O(\Delta t) + o(1).$$

The proof of first inequality then follows from the proof of Theorem 5.1 and (6.7). The second inequality also follows using

$$\mathbb{E}\left(\|e_t\|_2^2 \mid X_t = x\right) \leq \frac{1}{\Delta t}\sum_{m=1}^{d} |\sigma^{i,m}(x)|^2 + O(\Delta t^{\frac{-1}{2}}),$$

which can easily be derived using the Ito–Taylor expansions. ∎

6.1.3. General method for estimating drift. We have given methods which give second order estimates of α^i. To generate methods which give even higher order approximations, we note the similarities of the above methods to LMMs used in the numerical simulation of ODEs. Using the general LMM as a guide, we set up a general method for approximating μ^i:

$$\sum_{l=0}^{k} a_l \, \mu^i(X_{t_{n+l}}) \approx \sum_{l=1}^{p} b_l \, (X^i_{t_{n+l}} - X^i_{t_n}) \tag{6.10}$$

or

$$\left(\sum_{l=0}^{k} a_l \Theta_l\right) \alpha^i \approx \sum_{l=1}^{p} b_l D_l^i.$$

Keeping Remark 2 in mind, we can solve this using

$$\left(\sum_{l=0}^{k} a_l \Theta_0^* \Theta_l\right) \tilde{\alpha^i} = b_l \sum_{l=1}^{p} \Theta_0^* D_l^i. \tag{6.11}$$

The coefficients in (6.10) can be chosen to develop higher order methods. However, due to the stochastic nature of the problem, large amounts of data may be required to achieve the order in practice. We will need enough data to average over the randomness in the SDE, and the higher order methods can be sensitive to noise. More detailed investigation into the convergence of certain classes of methods for dynamics discovery can be found in [15] for deterministic systems.

6.2. Diffusion. In this section we will discuss improvements to the estimate for the diffusion. For some systems, particularly when the drift is large relative to the diffusion, the first order approximation given above may not be sufficient to obtain an accurate estimate of the diffusion coefficient. Using ideas similar to those in the previous section we can use the Ito–Taylor expansions to develop more accurate estimates of $\Sigma^{i,j}(X_t)$. However, these methods will be more complex; in addition to samples of X_t, some of these methods may also require data from the drift, $\mu^i(X_t)$ and $\mu^j(X_t)$.

6.2.1. Drift subtraction. Before discussing the higher order methods, we can make an improvement upon the first order method. In [27], Ragwitz and Kantz noted that by correcting for the effects of the drift in (4.2), we can make significant improvements to the estimate. To derive their estimate, we use the Ito–Taylor expansion for X_t, which gives us

$$X^i_{t+\Delta t} - X^i_t = \mu(X_t)\Delta t + \sum_{m=1}^{d} \sigma(X_t)\Delta W^m_t + R_t,$$

where $\Delta W_t = W_{t+\Delta t} - W_t$ is the increment of a d-dimensional Wiener process and R_t is the remainder term. This equation, with the remainder term excluded, actually gives the Euler–Marayama method for simulating SDEs. In essence, the approximation (4.2) uses

$$X^i_{t+\Delta t} - X^i_t \approx \sum_{m=1}^{d} \sigma^{i,m}(X_t)\Delta W^m_t$$

to approximate the increment of the Wiener process. However, (4.2) tosses out the $\mu(X_t)\Delta t$ term because it is of a higher order. If we include it, we get the more accurate

$$\sum_{m=1}^{d} \sigma^{i,m}\Delta W^m_t = (X^i_{t+\Delta t} - X^i_t) - \mu(X_t)\Delta t - R_t. \tag{6.12}$$

We can use this to generate a better approximation of $\Sigma^{i,j}$,

$$\Sigma^{i,j}(X_t) \approx \frac{(X^i_{t+\Delta t} - X^i_t - \mu^i(X_t)\Delta t)(X^j_{t+\Delta t} - X^j_t - \mu^j(X_t)\Delta t)}{2\Delta t}. \tag{6.13}$$

(We note that the estimate derived here is in slightly different form from that derived in [27], but will have a similar effect.) This approximation will be more accurate than (4.2), but it will be of same order with respect to Δt. Letting e_t be the error in (6.13), we can use the weak Ito–Taylor expansion to show

$$\mathbb{E}(e_t \,|\, X_t) = f(X_t)\frac{\Delta t}{2} + O(\Delta t^2), \qquad f = L^0\Sigma^{i,j} + \sum_{m=1}^{d}\left(\Sigma^{i,m}\frac{\partial \mu^j}{\partial x^m} + \Sigma^{j,m}\frac{\partial \mu^i}{\partial x^m}\right).$$

This gives an improvement over (5.7) by removing the $\mu^i\mu^j$ term in f (compared to Theorem 5.2). We note that this correction does not cancel all of the $O(\Delta t)$ terms in the error and thus does not improve the order of convergence. However, in systems where the drift dominates the diffusion the contributions of $\mu^i\mu^j$ will be large. For these systems, such as the Van der Pol (7.2) and Lorenz (7.3) examples presented in section 7, the improvement will be large. In systems where the drift is typically small, such as the system with a double well potential (7.1), the improvement will be modest.

In order to implement this method, we will need an approximation of μ^i. However, we can use the methods above to represent the drift as $\mu^i(X_t) \approx \theta(X_t)\tilde{\alpha}^i$. We can use this to set up the matrix equations

$$\Theta_0^*\Theta_0\tilde{\beta}^{i,j} = \frac{1}{\Delta t}(D_1^i - \Theta_0\tilde{\alpha}^i) \odot (D_1^j - \Theta_0\tilde{\alpha}^j) \tag{6.14}$$

and solve for $\tilde{\beta}^{i,j}$.

Remark 4. Equation (6.14) assumes that the same dictionary θ is used to estimate μ^i, μ^j, and $\Sigma^{i,j}$. In general, we could use separate dictionaries to estimate each of the parameters, since all we need are the approximations of the samples of $\mu^i(X_t)$ and $\mu^j(X_t)$ to estimate $\beta^{i,j}$.

6.3. Second order forward difference. While subtracting the drift from the differences $X^i_{t+\Delta t} - X^i_t$ gives marked improvements, we can also generate a higher order method using a two step forward difference, similar to the drift. The analysis for the estimation of the diffusion constant using the two step forward difference is essentially identical to that of the drift, so we will go through it briefly. Define the approximation

$$\Sigma^{i,j} \approx \frac{4(X^i_{t+\Delta t} - X^i_t)(X^j_{t+\Delta t} - X^j_t) - (X^i_{t+2\Delta t} - X^i_t)(X^j_{t+2\Delta t} - X^j_t)}{4\Delta t}. \tag{6.15}$$

As usual, letting e_t be the error in this approximation, we can use the Ito–Taylor expansions (3.4) and (3.8) to show that

$$\mathbb{E}(e_t) = O(\Delta t^2) \qquad \text{and} \qquad \mathbb{E}(|e_t|^2) = O(\Delta t).$$

This will gives us a second order method for the diffusion coefficients. We did not include the constants for the order Δt^2 for the sake of brevity, since the number of terms in the expressions can get quite large. We can use the approximation (6.15) to set up the matrix equations

$$\Theta_0^* \Theta_0 \tilde{\beta}^{i,j} = \frac{1}{4\Delta t} \Theta_0^* \left(4D_1^i \odot D_1^j - D_2^i \odot D_2^j \right), \tag{6.16}$$

which we can solve for $\tilde{\beta}^{i,j}$.

Theorem 6.3. *Consider the estimate $\tilde{\beta}^{i,j}$ given by solving (6.16). Then we have*

$$err_{mean} = O(\Delta t^2) + o(1)$$

and

$$err_{var} = \frac{1}{T} O(\Delta t) + o\left(\frac{1}{T}\right).$$

The proof of Theorem 6.3 is similar to the previous proofs. Additionally, we only give the leading order of the error, so deriving the bounds for $\mathbb{E}(e_t|X_t)$ and $\mathbb{E}(|e_t|^2|X_t)$ is simpler than the previous methods.

6.3.1. Trapezoidal method. Extending the trapezoidal approximation to estimating the diffusion coefficient is slightly trickier. Let $\Delta X^i_t = X^i_{t+\Delta t} - X^i_t$. If we attempt use the analogue to (6.6), we get

$$\Sigma^{i,j}(X_{t_{n+1}}) + \Sigma^{i,j}(X_t) = \frac{\Delta X^i_{t_n} \Delta X^j_{t_n}}{\Delta t} + R_{t_n},$$

with

$$\mathbb{E}(R_{t_n}) = f(X_{t_n})\Delta t + O(\Delta t^2), \qquad f = \mu^i \mu^j + \sum_{k=1}^{d} \left(\Sigma^{i,k} \frac{\partial \mu^j}{\partial x^k} + \Sigma^{j,k} \frac{\partial \mu^i}{\partial x^k} \right),$$

which is still only an order Δt method. However, we already demonstrated in (6.12) that correcting the difference ΔX_t^i for the drift can improve our approximation of $\sum_{m=1}^d \sigma^{i,m}\Delta W_t^m$. We will use the same trick here, except we will improve upon (6.12) by using the average values of μ^i and μ^j instead of the value at the left endpoint:

$$\sum_{m=1}^{d} \sigma^{i,m}\Delta W_t^m \approx (X_{t+\Delta t} - X_t) - \frac{\Delta t}{2}(\mu(X_t) + \mu(X_{t+\Delta t})).$$

If we use these differences to generate the trapezoidal method, we get

(6.17)
$$\Sigma^{i,j}(X_{t+\Delta t}) + \Sigma^{i,j}(X_t) \approx \frac{\left(\Delta X_t^i - \frac{\Delta t}{2}(\mu^i(X_t)+\mu^i(X_{t+\Delta t}))\right)\left(\Delta X_t^j - \frac{\Delta t}{2}(\mu^j(X_t) + \mu^j(X_{t+\Delta t}))\right)}{\Delta t}.$$

If we consider the error in (6.17), using the appropriate Ito–Taylor expansions we can show (see Appendix A.3)

$$|\mathbb{E}(e_t \,|\, X_t)| = O(\Delta t^2) \qquad \text{and} \qquad \mathbb{E}(|e_t|^2) = O(\Delta t).$$

Then, using the usual matrix notation, we can set up the equation

$$\Theta_0^*(\Theta_0 + \Theta_1)\tilde{\beta}^{i,j} = \frac{1}{\Delta t}\left(D_1^i - \frac{\Delta t}{2}(\Theta_0 + \Theta_1)\alpha^i\right) \odot \left(D_1^j - \frac{\Delta t}{2}(\Theta_0 + \Theta_1)\alpha^j\right). \tag{6.18}$$

We can solve this equation to get an order Δt^2 approximation of $\beta^{i,j}$.

Theorem 6.4. *Consider the estimate $\tilde{\beta}^{i,j}$ given by solving (6.18). Then we have*

$$err_{mean} = O(\Delta t^2) + o(1)$$

and

$$err_{var} = \frac{1}{T}O(\Delta t) + o\left(\frac{1}{T}\right).$$

The proof of Theorem 6.4 is similar to the previous proofs, using the appropriate error bounds. Although the order of the error is identical to that of Theorem 6.3, we will see that this method tends to have lower error. We did not include the constant terms for these errors for the sake of brevity, since the higher order Ito–Taylor expansions involve many terms.

7. Numerical examples.

In this section, we demonstrate the performance of the methods presented above on numerical examples. For each example, we will generate approximations $\tilde{\alpha}^i \approx \alpha^i$ and $\tilde{\beta}^{i,j} \approx \beta^{i,j}$. However, to present the data more simply, instead of computing the mean and mean squared error for each vector $\tilde{\alpha}^i$ and $\tilde{\beta}^{i,j}$, we will be aggregating the errors across all the coefficients. We will compute the mean error, normalized for the norms of α^i and $\beta^{i,j}$ using

$$Err_m = \left(\frac{\sum_{i=1}^d \|\mathbb{E}(\tilde{\alpha}^i) - \alpha^i\|_2^2}{\sum_{i=1}^d \|\alpha^i\|_2^2}\right)^{\frac{1}{2}} \quad \text{or} \quad Err_m = \left(\frac{\sum_{i\geq j\geq 1}^d \|\mathbb{E}(\tilde{\beta}^{i,j}) - \beta^{i,j}\|_2^2}{\sum_{i\geq j\geq 1}^d \|\beta^{i,j}\|_2^2}\right)^{\frac{1}{2}}.$$

Table 7.1
Summary of the methods for estimating the drift (μ^i) and the diffusion ($\Sigma^{i,j}$).

	Drift		Diffusion	
Name	Equation	Leading error term	Equation	Error
FD-Ord 1	(4.7)	$\frac{C_1}{2}\|L^0\mu^i\|_2\Delta t$	(4.8)	$O(\Delta t)$
FD-Ord 2	(6.4)	$\frac{2C_1}{3}\|(L^0)^2\mu^i\|_2\Delta t^2$	(6.16)	$O(\Delta t^2)$
Trapezoidal	(6.9)	$\frac{C_1}{12}\|(L^0)^2\mu^i\|_2\Delta t^2$	(6.18)	$O(\Delta t^2)$
Drift-Sub	-	-	(6.14)	$O(\Delta t)$

Similarly, we will calculate the normalized variance

$$Err_{var} = \frac{\sum_{i=1}^{d} Var\left(\tilde{\alpha}^i\right)}{\sum_{i=1}^{d} \|\alpha^i\|_2^2} \quad \text{or} \quad Err_{var} = \frac{\sum_{i\geq j\geq 1}^{d} Var\left(\tilde{\beta}^{i,j}\right)}{\sum_{i\geq j\geq 1}^{d} \|\beta^{i,j}\|_2^2}.$$

Since these errors are based on aggregating the errors for all of the components of α^i or $\beta^{i,j}$, they will demonstrate the same convergence rates as in Theorems 5.1, 5.2, and 6.1–6.4. The constants, however, may be different.

For each example, we will estimate the drift and diffusion using each of the methods described (see Table 7.1). The drift will be estimated using the first and second order forward differences, as well as the trapezoidal approximation. For the diffusion, we will use the first and second order forward differences, the drift-subtracted first order difference, and the trapezoidal method. For the drift-subtracted estimation, we will use the estimation for μ generated by the first order forward difference. Similarly, for the trapezoidal approximation for Σ, we will use the estimate generated by the trapezoidal approximation for μ.

7.1. Double well potential. Consider the SDE

$$dX_t = \left(-X_t^3 + \frac{1}{2}X_t\right)dX_t + \left(1 + \frac{1}{4}X_t^2\right)dW_t. \tag{7.1}$$

This equation represents a diffusion in the double well potential $U(x) = \frac{1}{4}x^4 - \frac{1}{2}x^2$. This example is similar to one considered in [1]. Without the diffusion, the trajectories of this system will settle toward one of two fixed points, depending on which basin of attraction it started in. With the stochastic forcing, a trajectory will move around in one basin of attraction until it gets sufficiently perturbed to move to the other basin. We also note that for the majority of the trajectory, the state will be near the point where the drift is zero, so the dynamics will be dominated by the diffusion. At these points, the trajectory will behave similarly to Brownian motion.

For the SINDy algorithm, we will use a dictionary of monomials in x up to degree 14:

$$\theta(x) = \begin{bmatrix} 1 & x & \cdots & x^{14} \end{bmatrix}.$$

This basis will be used to estimate both the drift and the diffusion. To generate the data for the algorithm, we simulated (7.1) using the Euler–Maruyama method 1,000 times with a time

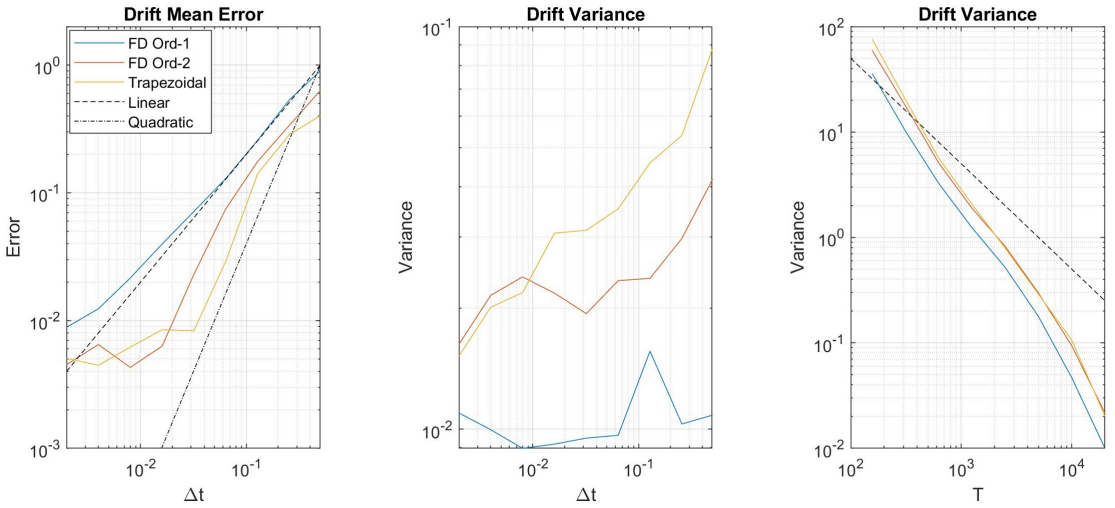

Figure 7.1. *(Left) The mean error in the estimation of the drift coefficients for the double well system (7.1) is plotted as a function of Δt. The error is approximated using 1,000 trajectories of length $T = 20{,}000$. (Center, right) The variance for each method is plotted against the sampling period, Δt, and the trajectory length, T. The trajectory length is fixed at $T = 20{,}000$ for the center plot, while the sampling period is fixed at $\Delta t = 0.004 = 4 \times 10^{-3}$ for the rightmost plot.*

step of 2×10^{-4} and a duration of 20,000. The initial condition was drawn randomly for each simulation from the standard normal distribution. The SINDy methods were then run on the data from each simulation for different sampling periods, Δt, and lengths of the trajectory, T. We use a minimum Δt of 0.002 so the simulation has a resolution of at least 10 steps between each data sample. The truncation parameters for the sparse solver were set at $\lambda = 0.005$ for the drift and $\lambda = 0.001$ for the diffusion.

As can be seen from Figure 7.1, the expected errors in all three methods for the drift were converging to zero as $\Delta t \to 0$. For small Δt, the expected estimate was within 1% of the true value. Additionally, the two higher order methods showed that, in expectation, they produce more accurate results and appear to converge more quickly, in line with Theorems 5.1, 6.1, and 6.2. For these methods, the expected error was as much as an order of magnitude smaller, depending on the size of Δt. The convergence rate for the first order method scales linearly with Δt, while the higher order methods appear to scale quadratically until $\Delta t = 0.02$, at which point there is likely not enough data to overcome the variance in the estimate.

The variance, however, is rather large relative to the size of the expected error for all three methods. This is likely due to the system tending to settle toward the points $x = \pm 1/\sqrt{2}$ where the drift is zero. Near these points, the dynamics are dominated by the diffusion, making it difficult to estimate the drift. As can be seen (noting the scale of the center plot), the variance does not change a great amount as Δt decreases, as is predicted for the estimates of the drift. As shown in the rightmost plot, the variance decreases as the length of the trajectory increases, slightly faster than linearly in $1/T$. In order to more fully benefit from using the higher order methods to the full extent, we would need a long enough trajectory to control the variance.

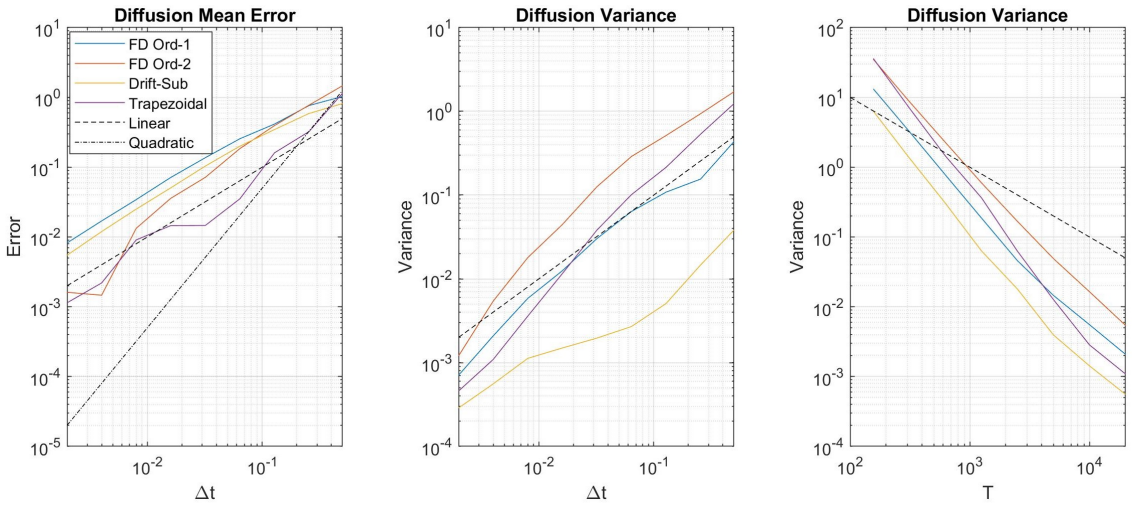

Figure 7.2. *(Left) The mean error in the estimation of the diffusion coefficients for the double well system* (7.1) *is plotted as a function of* Δt*. The error is approximated using* 1,000 *trajectories of length* $T = 20{,}000$*. (Center, right) The variance for each method is plotted against the sampling period,* Δt*, and the trajectory length,* T*. The trajectory length is fixed at* $T = 20{,}000$ *for the center plot, while the sampling period is fixed at* $\Delta t = 0.04 = 4 \times 10^{-3}$ *for the rightmost plot.*

For the diffusion, Figure 7.2 shows again that, as $\Delta t \to 0$, all of the methods do indeed converge in expectation. The Drift-Sub method slightly outperforms FD-Ord 1; the error is typically reduced by about 20%–30%. Of the two higher order methods, the trapezoidal method typically yields the best results, often an order of magnitude better than FD-Ord 1, although it does not appear to scale quadratically in Δt as predicted by the theorem. This is likely due to a lack of sufficient data to average over the noise. FD-Ord 2 also gives substantial improvements for small Δt. In contrast to the drift, the variance in the estimate of the diffusion does decrease as Δt goes to zero. The decrease appears to be proportional to Δt and slightly faster than linear in $1/T$, which is roughly in line with Theorems 5.2, 6.3, and 6.4.

7.2. Noisy Van der Pol oscillator. Consider the ODE

$$\begin{bmatrix} \dot{x}^1 \\ \dot{x}^2 \end{bmatrix} = \begin{bmatrix} x^2 \\ (1-(x^1)^2)x^2 - x^1 \end{bmatrix}.$$

This is the Van der Pol equation, which describes a nonlinear oscillator. We can perturb this equation by adding noise. We get the SDE

$$\begin{bmatrix} dX_t^1 \\ dX_t^2 \end{bmatrix} = \begin{bmatrix} X_t^2 \\ (1-(X_t^1)^2)X_t^2 - X_t^1 \end{bmatrix} dt + \sigma(X_t) dW_t, \tag{7.2}$$

where W_t is a two-dimensional Wiener process. For the simulations, we let

$$\sigma(x) = \frac{1}{2} \begin{bmatrix} 1 + 0.3x^2 & 0 \\ 0 & 0.5 + 0.2x^1 \end{bmatrix}.$$

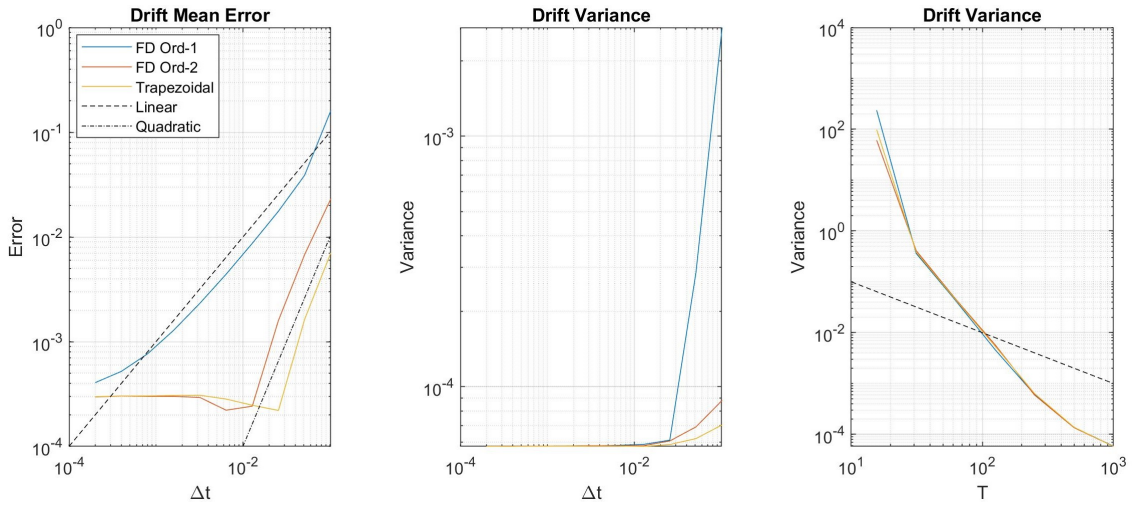

Figure 7.3. *(Left) The mean error in the estimation of the drift coefficients for the Van der Pol system (7.2) is plotted as a function of Δt. The error is approximated using 1,000 trajectories of length $T = 1{,}000$. (Center, right) The variance for each method is plotted against the sampling period, Δt, and the trajectory length, T. The trajectory length is fixed at $T = 1{,}000$ for the center plot, while the sampling period is fixed at $\Delta t = 0.008 = 8 \times 10^{-3}$ for the rightmost plot.*

We chose this system to represent a different type of limiting behavior, and the estimation of a stochastic Van der Pol oscillator was also considered in [2]. For this system, the dynamics settle around a limit cycle. While they will have a certain amount of randomness, the trajectories will demonstrate an approximately cyclic behavior. In particular, this also means that the drift will rarely be near zero, as opposed to the previous example where the drift was often small.

The dictionary we will use for the SINDy algorithm consists of all monomials in x^1 and x^2 up to degree 6:

$$\theta(x) = \begin{bmatrix} 1 & x^1 & x^2 & x^1x^2 & \cdots & (x^1)^2(x^2)^4 & x^1(x^2)^5 & (x^2)^6 \end{bmatrix}.$$

This basis will be used to estimate both the drift and diffusion. To generate the data for the algorithm, we simulated (7.2) using the Euler–Maruyama method 1,000 times with a time step of 2×10^{-5} and a duration of 1,000. Each component of the initial condition was drawn randomly for each simulation from the standard normal distribution. The SINDy methods were then run on the data from each simulation for different sampling periods, Δt, and lengths of the trajectory, T. As before, we use $\Delta t \geq 2 \times 10^{-4}$ to ensure that sampling period is at least 10 times the simulation time step. The truncation parameters for the sparse solver were set at $\lambda = 0.05$ for the drift and $\lambda = 0.02$ for the diffusion.

In Figure 7.3, we first note that the variance very quickly drops to about 5×10^{-5} and stays roughly constant as Δt decreases. This falls very much in line with Theorems 5.1, 6.1, and 6.2 which assert that the variance does not depend on the sample frequency, it only decreases with the trajectory length T. For the expected error, the FD-Ord 2 and trapezoidal methods show drastic improvements over FD-Ord 1, with the trapezoidal method reducing the error

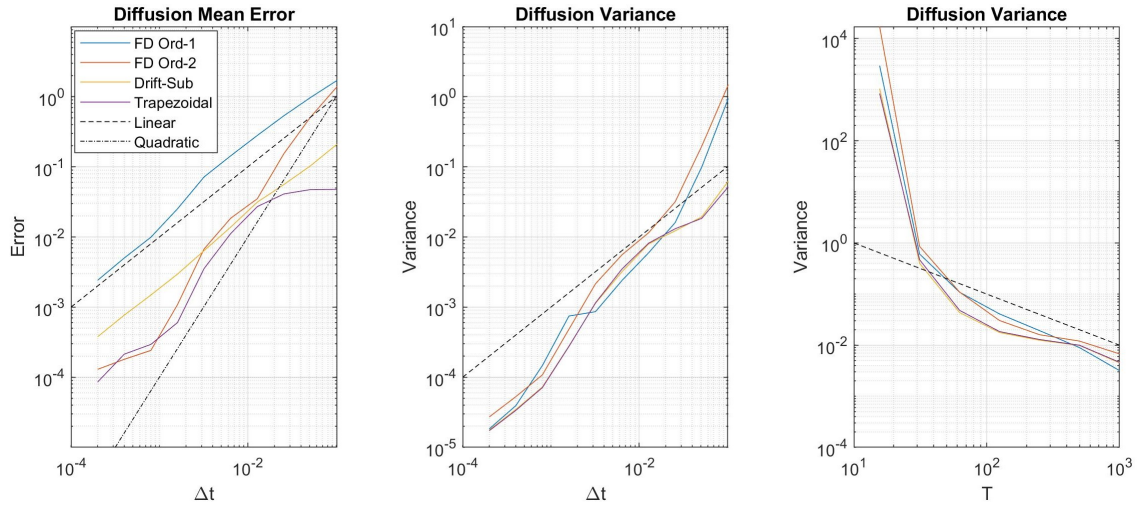

Figure 7.4. *(Left) The mean error in the estimation of the diffusion coefficients for the Van der Pol system* (7.2) *is plotted as a function of* Δt. *The error is approximated using* 1,000 *trajectories of length* $T = 1{,}000$. *(Center, right) The variance for each method is plotted against the sampling period,* Δt, *and the trajectory length,* T. *The trajectory length is fixed at* $T = 1{,}000$ *for the center plot, while the sampling period is fixed at* $\Delta t = 0.008 = 8 \times 10^{-3}$ *for the rightmost plot.*

by almost two orders of magnitude on some values of Δt. For the larger Δt, the slopes of the graphs demonstrate that these methods are converging at twice the order of the first order forward difference, as predicted by Theorems 5.1, 6.1, and 6.2. However, both second order methods quickly reach a point where the performance remained constant at about 2×10^{-4}. This is due to the lack of data to average over the random variation to sufficient precision. With sufficient data, we would expect the performance to continue to improve proportionally to Δt^2.

For the diffusion, Figure 7.4 demonstrates a greater separation in the performance of the different methods compared to the double well system. Here, the FD-Ord 1 and drift-subtracted methods both demonstrate the same first order convergence, as predicted in Theorem 5.2, but the drift-subtracted method demonstrates a substantially lower error, ranging from half an order to almost a full order of magnitude better. FD-Ord 2 begins at roughly the same error as FD-Ord 1 for large Δt, but convergences faster, as predicted by Theorem 6.3, until it gives more than an order of magnitude improvement for small Δt. Finally, although it is difficult to judge the speed of convergence for the trapezoidal method, it gives the most accurate results across all Δt. The variances for all of the methods behave similarly to the double well example and as expected, decreasing as $\Delta t \to 0$ and $T \to \infty$.

7.3. Noisy Lorenz attractor.

Consider the ODE

$$\dot{x} = \begin{bmatrix} \dot{x}^1 \\ \dot{x}^2 \\ \dot{x}^3 \end{bmatrix} = \begin{bmatrix} 10(x^2 - x^1) \\ x^1(28 - x^3) - x^2 \\ x^1 x^2 - \frac{8}{3} x^3 \end{bmatrix} = f(x).$$

This is the Lorenz system, which is famously a chaotic system exhibiting a strange attractor. If we perturb this equation by adding noise, we get the SDE

$$dX_t = f(X_t)dt + \sigma(X_t)dW_t, \tag{7.3}$$

where W_t is a three-dimensional Wiener process. The stochastic Lorenz process was also previous studied in the context of SDE identification in [12]. For this example, we let

$$\sigma(x) = \begin{bmatrix} 1+\sin(x^2) & 0 & \sin(x^1) \\ 0 & 1+\sin(x^3) & 0 \\ \sin(x^1) & 0 & 1-\sin(x^2) \end{bmatrix}.$$

To generate the data for the algorithm, we simulated (7.3) using the Euler–Maruyama method 1,000 times with a time step of 2×10^{-5} and a duration of 1,000. Each component of the initial condition was drawn randomly for each simulation from the standard normal distribution. The SINDy methods were then run on the data from each simulation for different sampling periods, Δt, and lengths of the trajectory, T. The truncation parameters for the sparse solver were set at $\lambda = 0.05$ for the drift and $\lambda = 0.02$ for the diffusion.

We will use different dictionaries to estimate the drift and diffusion. For the drift, the dictionary consists of all monomials in x^1, x^2, and x^3 up to degree 4:

$$\theta(x) = \begin{bmatrix} 1 & x^1 & x^2 & \cdots & x^1x^2(x^3)^3 & (x^2)^2(x^3)^3 & x^2(x^3)^4 & (x^3)^5 \end{bmatrix}.$$

As before, Figure 7.5 shows that the variance of the estimate for the drift decreases steadily as $T \to \infty$, while it approaches a minimum value as Δt decreases and remains constant

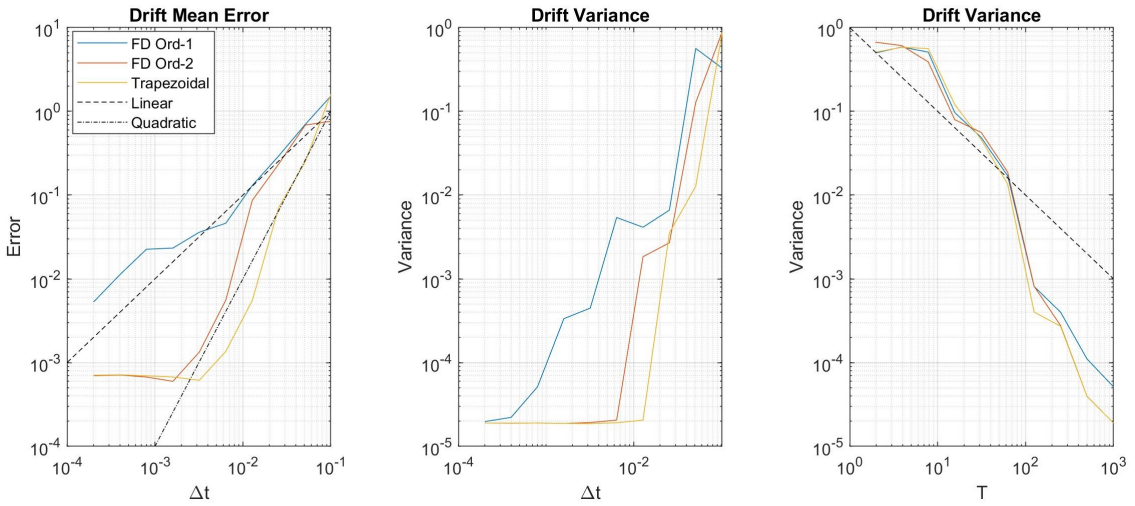

Figure 7.5. *(Left) The mean error in the estimation of the drift coefficients for the Lorenz system (7.3) is plotted as a function of Δt. The error is approximated using 1,000 trajectories of length $T = 1{,}000$. (Center, right) The variance for each method is plotted against the sampling period, Δt, and the trajectory length, T. The trajectory length is fixed at $T = 1{,}000$ for the center plot, while the sampling period is fixed at $\Delta t = 0.08 = 8 \times 10^{-2}$ for the rightmost plot.*

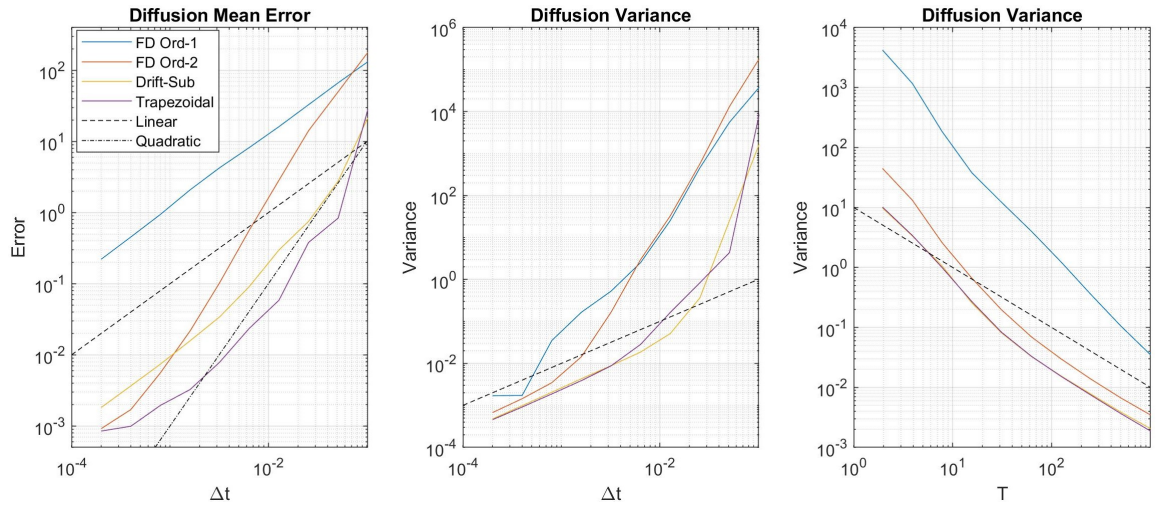

Figure 7.6. *(Left) The mean error in the estimation of the diffusion coefficients for the Lorenz system (7.3) is plotted as a function of Δt. The error is approximated using 1,000 trajectories of length $T = 1{,}000$. (Center, right) The variance for each method is plotted against the sampling period, Δt, and the trajectory length, T. The trajectory length is fixed at $T = 1{,}000$ for the center plot, while the sampling period is fixed at $\Delta t = 0.02 = 2 \times 10^{-2}$ for the rightmost plot.*

after reaching that minimum. In terms of the mean error, this example gives the clearest confirmation of the convergence rates demonstrated in Theorems 5.1, 6.1, and 6.2. The slopes of the plots show that the error with FD-Ord 1 is roughly proportional to Δt, while the FD-Ord 2 and trapezoidal methods converge at double the rate. For small Δt, the second order methods do not seem to improve, due to the lack of sufficient data to compute the averages to high enough precision.

To estimate the diffusion, we used a dictionary consisting of all monomials in $\sin(x^1)$, $\sin(x^2)$, and $\sin(x^3)$ up to degree 4:

$$\theta(x) = \begin{bmatrix} 1 & \sin(x^1) & \sin(x^2) & \cdots & \sin(x^1)\sin(x^2)\sin^2(x^3) & \sin(x^2)\sin^3(x^3) & \sin^4(x^3) \end{bmatrix}.$$

The error plot in Figure 7.6 provides the most compelling example of the improvements of the higher order methods for estimating the diffusion. FD-Ord 1 clearly demonstrates its order one convergence as $\Delta t \to 0$ (Theorem 5.2), but the error is quite large compared to the other methods. Even at our highest sampling frequency, $\Delta t = 2 \times 10^{-4}$, we only get slightly accurate results, with an error over 20%. For this system, the drift-subtracted method, although still first order, provides great improvements over FD-Ord 1, nearly two orders of magnitude better for most Δt. FD-Ord 2 also demonstrates the second order convergence given in Theorem 6.3, giving very accurate results for small Δt. Finally, the best performance again comes from the trapezoidal method, which gives the best performance across all Δt. As expected from Theorem 6.4, we can see that it converges faster than FD-Ord 1, but the convergence rate is not as clear as that of the other methods.

As for the variance, for all four methods it was roughly linear in $1/T$. It also decreased linearly with Δ_t once Δt was small enough to give accurate estimates overall. However, the trapezoidal and drift-subtracted methods both showed a substantially lower variance for larger Δt. This is likely because the drift tends to dominate the diffusion in this system. Both the drift-subtracted and trapezoidal methods correct for this, preventing the drift from having an effect on the estimate of the diffusion.

8. Measurement noise. As demonstrated in the numerical examples above, the estimates for the drift and diffusion can yield accurate results provided the sampling frequency is high enough and the length of the trajectory is long enough. However, all of the numerical examples presented assumed ideal data (i.e., there was no measurement noise in the observation of the state). For real systems, this is rarely the case.

For the estimates presented above, the errors introduced by nonideal data can be particularly large for high sampling frequencies due to the measurement noise being divided by Δt. The effects of noise on diffusion estimates have been studied especially in the context of single particle tracking [26, 24, 34]. Local estimates of the diffusion function which account for the noise have been presented [34, 12, 2], which can further be used to estimate the drift.

In this section, we will demonstrate how the methods presented above can be adapted to processes with measurement noise. For the drift, we will see that the approximations above can be directly adapted to handle noise using instrumental variables. For the diffusion, the approximation presented in [34] gives an unbiased estimate and can be extended by methods similar to those in section 6.2 for more accurate approximations.

For the duration of this section, we will assume that the measurement noise can be modeled as an independent and identically distributed (i.i.d.) Gaussian random vector with zero mean. Letting Y_t be the noisy measurement and δ_t be the noise, we have

$$Y_t = X_t + \delta_t, \qquad \delta_t \ \text{ i.i.d.}$$

Further, we will also assume the noise is small enough that we can evaluate our dictionary functions accurately. More precisely, for any dictionary function θ_k, we will assume

$$\theta_k(Y_t) = \theta(X_t + \delta_t) = \theta_k(X_t) + \sum_{i=1}^{d} (\nabla \theta_k)^T \delta_t + O(\|\delta_t\|^2),$$

and we can neglect the second order terms.

8.1. Stochastic force inference. In [12], the stochastic force inference (SFI) methodology estimates the drift and diffusion functions of an SDE with both ideal and noisy data. When using noisy data, SFI first estimates the diffusion using the local estimate in [34]. Then, to measure the drift, SFI approximates a Stratonovich integral, which is unbiased with noise, and uses the estimate of the diffusion to correct the Stratonovich integral to the Ito one.

Letting $\Delta Y_t^i = Y_{t+\Delta t}^i - Y_t$, SFI approximates the diffusion using

$$\Sigma^{i,j}(X_t) \approx \frac{(\Delta Y_t^i + \Delta Y_{t-\Delta t}^j)(\Delta Y_t^j + \Delta Y_{t-\Delta t}^j)}{4\Delta t} + \frac{\Delta Y_t^i \Delta Y_{t-\Delta t}^j + \Delta Y_{t-\Delta t}^i \Delta Y_t^j}{4\Delta t}. \tag{8.1}$$

This approximation gives an estimate that converges with order Δt. Using the notation of (6.1) with the matrices populated using the noisy data Y_t, we can set up the normal equation to solve for $\tilde{\beta}^{i,j}$ as

$$\Theta_1^*\Theta_1\tilde{\beta^{i,j}} = \frac{1}{4\Delta t}\Theta_1^*\left[D_2^i \odot D_2^j + D_1^i \odot (D_2^j - D_1^j) + (D_2^i - D_1^i) \odot D_2^j\right]. \tag{8.2}$$

When estimating the drift, errors arise from the interaction of the noise in the approximation of μ and the effects of the noise on the dictionary function. To combat this, SFI evaluates a discretized Stratonovich integral and uses an estimate of the diffusion to convert the Stratonovich integral to an Ito one. The symmetry of the Stratonovich integral removes the bias introduced by the noise. The normal equations to summarize this method are

$$\Theta_0^*\Theta_0\tilde{\alpha^i} = \frac{1}{2\Delta t}(\Theta_0 + \Theta_1)^* D_1^i + \sum_{x=0}^{N-1}\sum_{j=1}^{d}\Sigma^{i,j}(X_{t_n})\frac{\partial\theta}{\partial x^j}(X_{t_n}). \tag{8.3}$$

For the evaluation $\Sigma^{i,j}(X_t)$ in this equation we can use the approximation (8.1). The first term on the right-hand side of this equation approximates the Stratonovich integral $\int_0^T \theta \circ dX_t$, while the second term corrects it to the Ito integral. For more detailed analysis of these methods, see [12]. While this estimate is unbiased, it does have the disadvantages of requiring the differential of θ and using an estimate of Σ, which will also have some error.

8.2. Instrumental variables for estimating drift. While the SFI method is unbiased, it does have the disadvantages of using an estimate of Σ, which will also have some error, and requiring knowledge of the differential of θ. However, we can adapt the estimates in section 6.1 to be unbiased. These methods will realize the same order of convergence (with respect to Δt) as the methods with ideal data in the large data limit. Additionally, they have the advantage of being simple to implement and do not require the differential of θ.

Consider the first order forward difference in the presence of noise

$$\mu(X_t) \approx \frac{Y_{t+\Delta t} - Y_t}{\Delta t} = \frac{X_{t+\Delta t} - X_t}{\Delta t} + \frac{\delta_{t+\Delta t} - \delta t}{\Delta t}.$$

Since this approximation is linear in the noise δ_t, its expected value is unaffected by the noise. All of the difference methods presented in section 6.1 have this property, since they are linear. The bias only comes from the interaction of the noise in the numerical derivatives and the noise in the dictionary. The bias is given by

$$\mathbb{E}\left(\theta_k(Y_t)\frac{\delta_{t+\Delta t} - \delta_t}{\Delta t}\right) \approx \mathbb{E}\left[\left(\theta_k(X) + (\nabla\theta_k)^T\delta_t\right)\left(\frac{\delta_{t+\Delta t} - \delta_t}{\Delta t}\right)\right] = Cov(\delta_t)\nabla\theta_k,$$

where $Cov(\delta_t) = \mathbb{E}(\delta_t\delta_t^T)$ is the covariance matrix of δ_t. However, if we use the previous dictionary values, $\theta(Y_{t-\Delta t})$, we have

$$\mathbb{E}\left(\theta(Y_{t-\Delta t})\frac{Y_{t+\Delta_t} - Y_t}{\Delta_t}\right) = \theta(X_{t-\Delta t})\frac{X_{t+\Delta t} - X_t}{\Delta t}$$

since the realizations of the noise in $Y_{t-\Delta t}, Y_t$, and $Y_{t+\Delta t}$ are all independent. This amounts to using $\Theta(Y_{t-\Delta t})$ as a set of instrumental variables (see [32]). The normal equation for this regression is

$$\Theta_0^* \Theta_1 \tilde{\alpha}^i = \frac{1}{\Delta t} \Theta_0 D_1^i. \tag{8.4}$$

Similar to the first order method above, we can find a set of normal equations for the trapezoidal method for drift using instrumental variables:

$$\frac{1}{2} \Theta_0 \left(\Theta_1 + \Theta_2\right) \tilde{\alpha}^i = \frac{1}{\Delta t} \Theta_0 D_1^i. \tag{8.5}$$

8.3. Improving the diffusion estimate?. Equation (8.1) gives an $O(\Delta t)$ approximation of $\Sigma^{i,j}(X_t)$ in expectation. We can improve this estimate in a similar manner to the trapezoidal method for diffusion. Let

$$s^i(t) = Y_{t+\Delta t}^i - Y_t^i - \frac{\Delta t}{2}\left(\mu^i(Y_t) + \mu^i(Y_{t+\Delta t})\right), \quad q^i(t) = Y_{t+2\Delta t}^i - Y_t^i - \Delta t\left(\mu^i(Y_t) + \mu^i(Y_{t+2\Delta t})\right).$$

Then we can use the approximation

$$\Sigma^{i,j}(X_t) + \Sigma^{i,j}(X_{t+2\Delta t}) \approx \frac{1}{2\Delta t}\left(q^i(t)q^j(t) + s^i(t)s^i(t+1) + s^i(t+1)s^i(t)\right). \tag{8.6}$$

Letting

$$S_n^i = \begin{bmatrix} s^i(t_n) & s^i(t_{n+1}) & \cdots s^i(t_{N+n-1}) \end{bmatrix}^T \quad \text{and} \quad Q_n^i = \begin{bmatrix} q^i(t_n) & q^i(t_{n+1}) & \cdots q^i(t_{N+n-1}) \end{bmatrix}^T,$$

we can set up the instrumental variables regression

$$\Theta_0^*(\Theta_1 + \Theta_3)\tilde{\beta}^{i,j} = \frac{1}{2\Delta t}\Theta_0^*\left(Q_1^i \odot Q_1^j + S_1^i \odot S_2^j + S_2^i \odot S_1^j\right) \tag{8.7}$$

to solve for $\tilde{\beta}^{i,j}$.

8.4. Van der Pol oscillator. We now consider the stochastic Van der Pol oscillator given by (7.2) with measurement noise on the state X_t. Each component of the noise δ_t^i is drawn from a normal distribution with zero mean and a standard deviation of 0.02. The system was simulated 1,000 times with a time step of 4×10^{-5} and a duration of 1,000. As before, we use $\Delta t \geq 4 \times 10^{-4}$ to ensure that the sampling period is at least 10 times the simulation time step. The truncation parameters for the sparse solver were set at $\lambda = 0.05$ for both the drift and the diffusion. Using this data, we test the noise-corrected methods presented in this section along with the first order and trapezoidal methods tested in section 7.

As can be seen from the error plots, while the methods which don't account for noise may be somewhat accurate estimates for large Δt, as $\Delta t \to 0$ they diverge from the true values of α^i and $\beta^{i,j}$. For the drift estimate (see Figure 8.1), the SFI and methods using instrumental variables improve as $\Delta t \to 0$. The instrumental variables methods, however, tend to give more accurate results, with the trapezoidal IV method greatly outperforming the others for larger Δt. For the diffusion (see Figure 8.2), the estimate in SFI converges toward the true values of $\beta^{i,j}$ as $\Delta_t \to 0$. The trapezoidal method, however, reaches the same level of accuracy for much larger Δt.

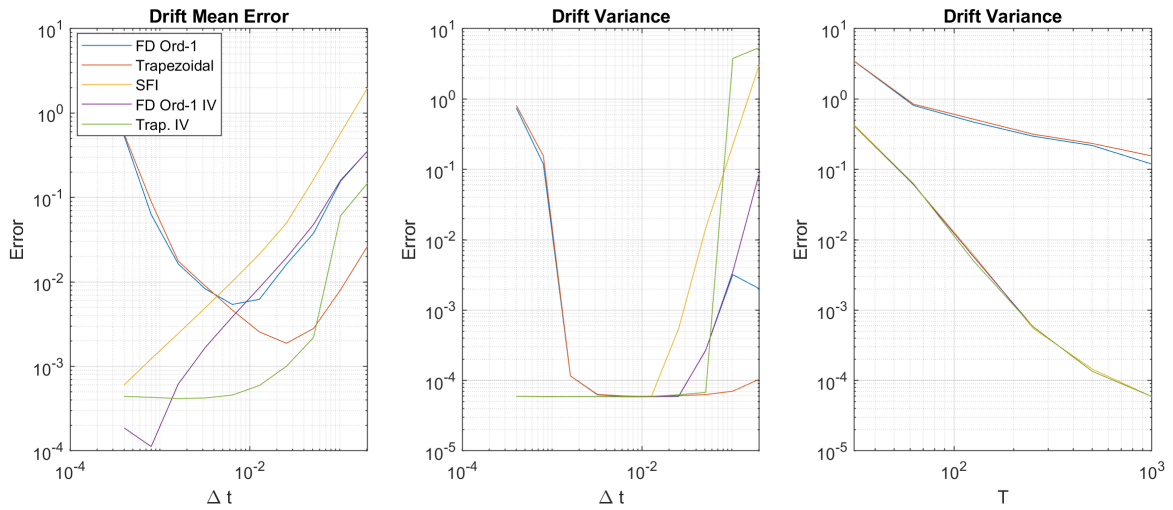

Figure 8.1. *(Left) The mean error in the estimation of the drift coefficients for the Van der Pol system* (7.2) *with measurement noise is plotted as a function of* Δt*. The error is approximated using* 1,000 *trajectories of length* $T = 1{,}000$*. (Center, right) The variance for each method is plotted against the sampling period,* Δt*, and the trajectory length,* T*. The trajectory length is fixed at* $T = 1{,}000$ *for the center plot, while the sampling period is fixed at* $\Delta t = 0.008 = 8 \times 10^{-3}$ *for the rightmost plot.*

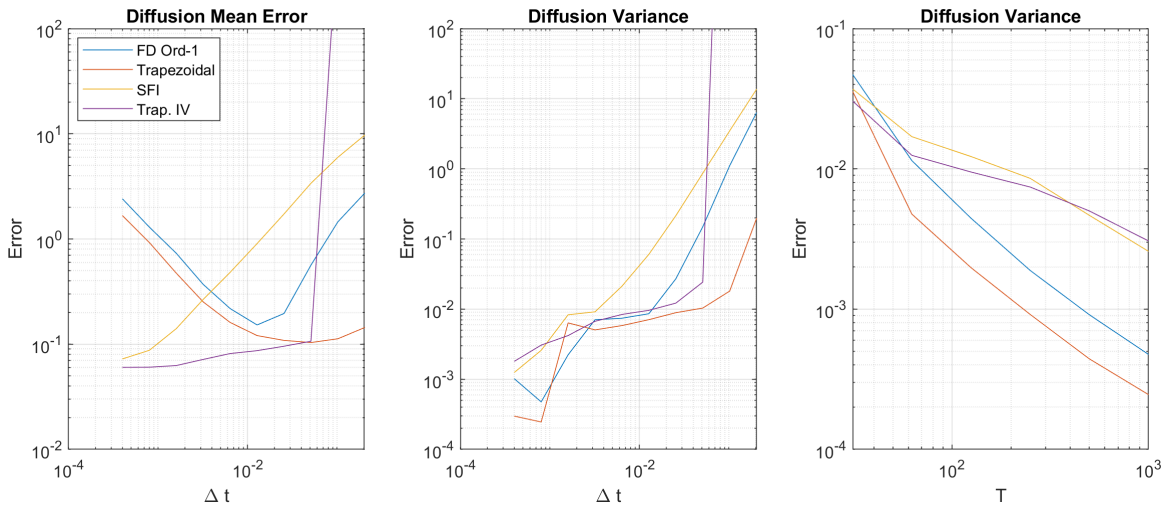

Figure 8.2. *(Left) The mean error in the estimation of the diffusion coefficients for the Van der Pol system* (7.2) *with measurement noise is plotted as a function of* Δt*. The error is approximated using* 1,000 *trajectories of length* $T = 1{,}000$*. (Center, right) The variance for each method is plotted against the sampling period,* Δt*, and the trajectory length,* T*. The trajectory length is fixed at* $T = 1{,}000$ *for the center plot, while the sampling period is fixed at* $\Delta t = 0.008 = 8 \times 10^{-3}$ *for the rightmost plot.*

9. Conclusion. As was shown in this and previous papers [1, 8, 5], the SINDy algorithm can be used to accurately estimate the parameters of a stochastic differential equation. However, the significant amount of noise involved requires one to use either a great deal of data

(i.e., a long time series) and/or methods which improve the robustness of SINDy to noise. Unfortunately, even if SINDy should identify all of the correct dictionary functions present in the dynamics, we showed that the sampling frequency limits the accuracy of the results when using the first order Kramers–Moyal formulas to estimate the drift and diffusion. The necessity for high sampling frequencies, combined with long trajectories, makes SINDy a data hungry algorithm.

The higher order estimates presented in this paper allow us to overcome the $O(\Delta t)$ convergence given in [1, 12]. With the higher order methods we can compute accurate estimations of the SDEs using far lower sampling frequencies. In addition to making SINDy a more accurate system identification tool, these improvements also greatly reduce the data requirements to feed the algorithm. By achieving accurate results at lower sampling frequencies we can reduce the data acquisition constraint, which makes SINDy a more feasible system identification method for SDEs.

Appendix A. Error derivations for section 5.

In Theorems 5.1–5.2 and 6.1–6.4, we used estimates of the drift and diffusion based on finite differences. Most of the derivations are straightforward and follow almost immediately from the Ito–Taylor expansions. However, bounding the estimate for the variance in the second order difference for the drift and bounding the first order estimate for the diffusion require a little extra work, so we include them here.

A.1. Drift: Second order forward difference.

The error in the second order forward difference estimate (6.2) for the drift is given by

$$e_{t_n} = \mu^i(X_{t_n}) - \frac{-3X^i_{t_n} + 4X^i_{t_{n+1}} - X^i_{t_{n+2}}}{2\Delta t}.$$

Using (3.5) in the estimate above gives us

$$\mathbb{E}(e_t \,|\, X_t) = -\frac{\Delta t}{3}(L^0)^2\mu^i + O(\Delta t^3). \tag{A.1}$$

Proof of Theorem 6.1. The proof of the estimate on the mean error follows from (A.1) and the proof of Theorem 5.1. Now, Let

$$\Theta_0 = \begin{bmatrix} \theta(X_{t_0}) \\ \theta(X_{t_1}) \\ \vdots \\ \theta(X_{t_{N-1}}) \end{bmatrix} \quad \text{and} \quad E = \begin{bmatrix} e_0 \\ e_1 \\ \vdots \\ e_{N-1} \end{bmatrix}.$$

To estimate the variance, we need to find $\mathbb{E}(\|\frac{1}{N}\Theta_0^* E\|_2^2)$. To do this, we will use the strong expansion (3.7) and obtain

$$e_t = \frac{1}{2\Delta t}\left(\sum_{m=1}^{d} \sigma^{i,m}(3\Delta W_t^m - \Delta W_{t+1}^m) + R_t\right)$$

with $\mathbb{E}(|R_t|^2) = O(\Delta t^2)$. Then we use

$$\begin{aligned}\frac{1}{N}\Theta_0^* E &= \frac{1}{N}\sum_{n=0}^{N-1}\theta^*(X_{t_n})e_{t_n} = \frac{1}{N}\sum_{n=0}^{N-1}\theta^*(X_{t_n})\left(\sum_{m=1}^{d}\sigma^{i,m}(X_{t_n})\frac{3\Delta W_{t_n}^m - \Delta W_{t_{n+1}}^m}{2\Delta t} + \frac{R_{t_n}}{2\Delta t}\right)\\ &= \frac{1}{2T}\sum_{n=0}^{N-1}\theta^*(X_{t_n})\left(\sum_{m=1}^{d}(3\sigma^{i,m}(X_{t_n}) - \sigma^{i,m}(X_{t_{n-1}}))\Delta W_{t_n}^m + R_{t_n}\right) + R_1\\ &= \frac{1}{T}\sum_{n=0}^{N-1}\theta^*(X_{t_n})\left(\sum_{m=1}^{d}(\sigma^{i,m}(X_{t_n}) + R_{t_n}^m)\Delta W_{t_n}^m + R_{t_n}\right) + R_1,\end{aligned}$$

where

$$R_1 = \frac{1}{T}\sum_{m=1}^{d}\left(\theta^*(X_{t_0})\sigma^{i,m}(X_{t_0})\Delta W_{t_0}^m - \theta^*(X_{t_N})\sigma^{i,m}(X_{t_N})\Delta W_{t_N}^m\right)$$

and $\mathbb{E}(|R_{t_n}^m|^2) = O(\Delta t)$. The second line comes from rearranging the indices of the sum, which gives the remainder R_1, and the last line uses the Ito–Taylor expansion of $\sigma^{i,m}$, which gives the remainder R_2. Combining all of the errors gives us

$$\frac{1}{N}\Theta_0^* E = \frac{1}{T}\sum_{n=0}^{N-1}\sum_{m=1}^{d}\theta^*(X_{t_n})\sigma^{i,m}(X_{t_n})\Delta W_{t_n}^m + R$$

with $\mathbb{E}(R^2) = O(\Delta t^2)$. Taking the expectance of the square of this last equation gives us

$$\mathbb{E}\left(\left\|\frac{1}{N}\Theta_0^* E\right\|_2^2\right) \le \frac{1}{T^2}\sum_{n=0}^{N}\sum_{m=1}^{d}\|\theta^*(X_{t_n})\|_2^2\sigma^{i,m}(X_{t_n})^2\Delta t + O(\Delta t^{\frac{3}{2}}).$$

Using this, the rest of the proof follows that of Theorem 5.1. ∎

A.2. Diffusion: First order forward difference. For Theorem 5.2, we use (4.2) to approximate the diffusion matrix elements. We need to bound the errors to give (5.7) and (5.8) for the proof. From the approximation (4.2), the error is

$$e_t = \frac{(X_{t+\Delta t}^i - X_t^i)(X_{t+\Delta t}^j - X_t^j)}{2\Delta t} - \Sigma^{i,j}(X_t).$$

The expected error, $\mathbb{E}(e_t|X_t)$, is easy to bound using (3.6). To calculate the squared error, $\mathbb{E}(|e_t|^2|X_t)$, we will use the strong expansion (3.8). This gives us

$$e_t = \frac{1}{2\Delta t}\left(\sum_{k,l=1}^{d}(\sigma^{k,i}\sigma^{l,j}(X_t) + \sigma^{k,j}\sigma^{l,i}(X_t))I_{i,j} + R_t\right) \tag{A.2}$$

with $\mathbb{E}(|R_t|^2|X_t) = O(\Delta t^3)$. From Lemma 5.7.2 of [17], we have

$$\mathbb{E}(I_{(k,l)}I_{(m,n)}) = \begin{cases}0, & (k,l) \neq (m,n),\\ \frac{\Delta t^2}{2}, & k = m, l = n.\end{cases}$$

Then, squaring (A.2), we get

$$\begin{aligned}\mathbb{E}(|e_t|^2 \,|\, X_t) &= \frac{1}{4\Delta t^2}\sum_{k,l=1}^{d}(\sigma^{k,i}\sigma^{l,j}(X_t)+\sigma^{k,j}\sigma^{l,i}(X_t))^2\frac{\Delta t^2}{2}+O(\Delta t^{\frac{1}{2}})\\ &= \frac{1}{8}\sum_{k,l=1}^{d}2\left((\sigma^{k,i}\sigma^{l,j}(X_t))^2+\sigma^{k,i}\sigma^{l,j}\sigma^{k,j}\sigma^{l,i}(X_t)\right)+O(\Delta t^{\frac{1}{2}})\\ &= \Sigma^{i,i}(X_t)\Sigma^{j,j}(X_t)+\Sigma^{i,j}(X_t)^2+O(\Delta t^{\frac{1}{2}}),\end{aligned}$$

which gives us (5.8).

A.3. Trapezoidal approximation for diffusion. The trapezoidal method to approximate the diffusion is given by (6.17):

(A.3)
$$\Sigma^{i,j}(X_{t+\Delta t})+\Sigma^{i,j}(X_t)\approx\frac{\left(\Delta X_t^i-\frac{\Delta t}{2}(\mu^i(X_t)+\mu^i(X_{t+\Delta t}))\right)\left(\Delta X_t^j-\frac{\Delta t}{2}(\mu^j(X_t)+\mu^j(X_{t+\Delta t}))\right)}{\Delta t}.$$

We claim that the error in this approximation can be bounded by

$$|\mathbb{E}(e_t \,|\, X_t)| = O(\Delta t^2) \qquad \text{and} \qquad \mathbb{E}(|e_t|^2) = O(\Delta t).$$

To achieve this, we let $\Delta\mu_t^i = \mu^i(X_{t+\Delta t}) - \mu^i(X_t)$ and rewrite the right-hand side of (A.3) as

(A.4)
$$\frac{\left(\Delta X_t^i-\mu^i(X_t)\Delta t-\frac{\Delta t}{2}\Delta\mu^i\right)\left(\Delta X_t^j-\mu^j(X_t)\Delta t-\frac{\Delta t}{2}\Delta\mu_t^j\right)}{\Delta t}=\Sigma^{i,j}(X_t)+\Sigma^{i,j}(X_{t+\Delta t})+e_t.$$

We will look at several of the cross terms separately on the left-hand side. Using (3.6) and (3.5), we can see that the first term will be

$$\mathbb{E}\left(\frac{\left(\Delta X_t^i-\mu^i(X_t)\Delta t\right)\left(\Delta X_t^j-\mu^j(X_t)\Delta t\right)}{\Delta t}\Big| X_t\right)=2\Sigma^{i,j}(X_t)+h(X_t)\Delta t+O(\Delta t^2),$$

where

$$h = L^0\Sigma^{i,j}+\sum_{k=1}^{d}\left(\Sigma^{j,k}\frac{\partial\mu^i}{\partial x^k}+\Sigma^{i,k}\frac{\partial\mu^j}{\partial x^k}\right).$$

Next we will consider the $\Delta X_t^i\Delta\mu_t^j$ terms. If we use the weak Ito–Taylor expansion of $f(x) = (x^i - X_t^i)(\mu^j(x) - \mu^j(X_t))$, holding X_t fixed, we see that

$$\mathbb{E}\left(\Delta X_t^i\Delta\mu_t^j|X_t\right)=\mathbb{E}(f(X_{t+\Delta t})|X_t)=L^0f(X_t)\Delta t+O(\Delta t^2)=2\sum_{k=1}^{d}\Sigma^{i,k}\frac{\partial\mu^j}{\partial x^k}.$$

These will cancel the last terms in h. All other terms on the left-hand side will be of higher order. For the right-hand side of (A.4), we can use the Ito–Taylor expansion of $\Sigma^{i,j}$ to show

$$\mathbb{E}\left(\Sigma^{i,j}(X_{t+\Delta t})|X_t\right) = \Sigma^{i,j}(X_t) + L^0\Sigma^{i,j}\Delta t + O(\Delta t^2).$$

Combining all of these together, we see that

$$\Sigma^{i,j}(X_t) + \Sigma^{i,j}(X_{t+\Delta t}) = \frac{\left(\Delta X_t^i - \mu^i(X_t)\Delta t - \frac{\Delta t}{2}\Delta\mu^i\right)\left(\Delta X_t^j - \mu^j(X_t)\Delta t - \frac{\Delta t}{2}\Delta\mu_t^j\right)}{\Delta t} + e_t$$

with $\mathbb{E}(e_t|x_t) = O(\Delta t^2)$. The bound for the squared error, $\mathbb{E}(|e_t|^2 \,|\, X_t) = O(\Delta t)$, follows easily from (3.8), since the correction terms added are all of higher order.

REFERENCES

[1] L. Boninsegna, F. Nüske, and C. Clementi, *Sparse learning of stochastic dynamical equations*, J. Chem. Phys., 148 (2018), 241723.

[2] D. B. Brückner, P. Ronceray, and C. P. Broedersz, *Inferring the dynamics of underdamped stochastic systems*, Phys. Rev. Lett., 125 (2020), 058103.

[3] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proc. Natl. Acad. Sci. USA, 113 (2016), pp. 3932–3937.

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Sparse identification of nonlinear dynamics with control (SINDYc)*, IFAC-PapersOnLine, 49 (2016), pp. 710–715.

[5] J. L. Callaham, J.-C. Loiseau, G. Rigas, and S. L. Brunton, *Nonlinear stochastic modelling with Langevin regression*, Proc. A, 477 (2021), 20210092.

[6] X. Chen and I. Timofeyev, *Non-parametric estimation of stochastic differential equations from stationary time-series*, J. Stat. Phys., 186 (2022), 1.

[7] F. Comte, V. Genon-Catalot, and Y. Rozenholc, *Penalized nonparametric mean square estimation of the coefficients of diffusion processes*, Bernoulli, 13 (2007), pp. 514–543.

[8] M. Dai, T. Gao, Y. Lu, Y. Zheng, and J. Duan, *Detecting the maximum likelihood transition path from data of stochastic dynamical systems*, Chaos, 30 (2020), 113124.

[9] U. Fasel, J. N. Kutz, B. W. Brunton, and S. L. Brunton, *Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control*, Proc. A, 478 (2022), 20210904.

[10] R. Friedrich, J. Peinke, M. Sahimi, and M. R. R. Tabar, *Approaching complexity by stochastic methods: From biological systems to turbulence*, Phys. Rep., 506 (2011), pp. 87–162.

[11] R. Friedrich, S. Siegert, J. Peinke, St. Lück, M. Siefert, M. Lindemann, J. Raethjen, G. Deuschl, and G. Pfister, *Extracting model equations from experimental data*, Phys. Lett. A, 271 (2000), pp. 217–222.

[12] A. Frishman and P. Ronceray, *Learning force fields from stochastic trajectories*, Phys. Rev. X, 10 (2020), 021009.

[13] K. Kaheman, J. N. Kutz, and S. L. Brunton, *SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics*, Proc. A, 476 (2020), 20200279.

[14] E. Kaiser, J. N. Kutz, and S. L. Brunton, *Sparse identification of nonlinear dynamics for model predictive control in the low-data limit*, Proc. A, 474 (2018), 20180335.

[15] R. T. Keller and Q. Du, *Discovery of dynamics using linear multistep methods*, SIAM J. Numer. Anal., 59 (2021), pp. 429–455, https://doi.org/10.1137/19M130981X.

[16] R. Khasminskii, *Stochastic Stability of Differential Equations*, Stoch. Model. Appl. Probab. 66, Springer-Verlag, Berlin, Heidelberg, 2011.

[17] P. E. Kloeden and E. Platen, *Stochastic differential equations*, in Numerical Solution of Stochastic Differential Equations, Springer, 1992, pp. 103–160.

[18] E. B. Kosmatopoulos, M. M. Polycarpou, M. A. Christodoulou, and P. A. Ioannou, *High-order neural network structures for identification of dynamical systems*, IEEE Trans. Neural Netw., 6 (1995), pp. 422–431.
[19] L. Ljung, *System Identification: Theory for the User*, Pearson Education, 1998.
[20] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Inferring biological networks by sparse identification of nonlinear dynamics*, IEEE Trans. Mol. Biol. Multi-Scale Commun., 2 (2016), pp. 52–63.
[21] D. A. Messenger and D. M. Bortz, *Weak SINDy for partial differential equations*, J. Comput. Phys., 443 (2021), 110525.
[22] D. A. Messenger and D. M. Bortz, *Weak SINDy: Galerkin-based data-driven model selection*, Multiscale Model. Simul., 19 (2021), pp. 1474–1497, https://doi.org/10.1137/20M1343166.
[23] I. Mezić, *Spectral properties of dynamical systems, model reduction and decompositions*, Nonlinear Dynam., 41 (2005), pp. 309–325.
[24] X. Michalet and A. J. Berglund, *Optimal diffusion coefficient estimation in single-particle tracking*, Phys. Rev. E, 85 (2012), 061916.
[25] K. S. Narendra and K. Parthasarathy, *Identification and control of dynamical systems using neural networks*, IEEE Trans. Neural Netw., 1 (1990), pp. 4–27.
[26] H. Qian, M. P. Sheetz, and E. L. Elson, *Single particle tracking. Analysis of diffusion and flow in two-dimensional systems*, Biophys. J., 60 (1991), pp. 910–921.
[27] M. Ragwitz and H. Kantz, *Indispensable finite time corrections for Fokker-Planck equations from time series data*, Phys. Rev. Lett., 87 (2001), 254501.
[28] O. Reiersøl, *Confluence Analysis by Means of Instrumental Sets of Variables*, Ph.D. thesis, Almqvist & Wiksell, 1945.
[29] P. J. Schmid, *Dynamic mode decomposition of numerical and experimental data*, J. Fluid Mech., 656 (2010), pp. 5–28.
[30] F. Sicard, V. Koskin, A. Annibale, and E. Rosta, *Position-dependent diffusion from biased simulations and Markov state model analysis*, J. Chem. Theory Comput., 17 (2021), pp. 2022–2033.
[31] S. Siegert, R. Friedrich, and J. Peinke, *Analysis of data sets of stochastic systems*, Phys. Lett. A, 243 (1998), pp. 275–280.
[32] T. Söderström and P. Stoica, *Instrumental Variable Methods for System Identification*, Lect. Notes Control Inf. Sci. 57, Springer-Verlag, Berlin, New York, 1983.
[33] R. Tibshirani, *Regression shrinkage and selection via the lasso*, J. Roy. Statist. Soc. Ser. B, 58 (1996), pp. 267–288.
[34] C. L. Vestergaard, P. C. Blainey, and H. Flyvbjerg, *Optimal estimation of diffusion coefficients from single-particle trajectories*, Phys. Rev. E, 89 (2014), 022726.
[35] M. Wanner and I. Mezić, *Robust approximation of the stochastic Koopman operator*, SIAM J. Appl. Dyn. Syst., 21 (2022), pp. 1930–1951, https://doi.org/10.1137/21M1414425.
[36] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, *A data–driven approximation of the Koopman operator: Extending dynamic mode decomposition*, J. Nonlinear Sci., 25 (2015), pp. 1307–1346.
[37] L. Zhang and H. Schaeffer, *On the convergence of the SINDy algorithm*, Multiscale Model. Simul., 17 (2019), pp. 948–972, https://doi.org/10.1137/18M1189828.