Rapid Communications

Using noisy or incomplete data to discover models of spatiotemporal dynamics

Patrick A. K. Reinbold, Daniel R. Gurevich, and Roman O. Grigoriev
School of Physics, Georgia Institute of Technology, Atlanta, Georgia 30332-0430, USA

(Received 30 September 2019; published 22 January 2020)

Sparse regression has recently emerged as an attractive approach for discovering models of spatiotemporally complex dynamics directly from data. In many instances, such models are in the form of nonlinear partial differential equations (PDEs); hence sparse regression typically requires the evaluation of various partial derivatives. However, accurate evaluation of derivatives, especially of high order, is infeasible when the data are noisy, which has a dramatic negative effect on the result of regression. We present an alternative and rather general approach that addresses this difficulty by using a weak formulation of the problem. For instance, it allows accurate reconstruction of PDEs involving high-order derivatives, such as the Kuramoto-Sivashinsky equation, from data with a considerable amount of noise. The flexibility of our approach also allows reconstruction of PDE models that involve latent variables which cannot be measured directly with acceptable accuracy. This is illustrated by reconstructing a model for a weakly turbulent flow in a thin fluid layer, where neither the forcing nor the pressure field is known.

DOI: 10.1103/PhysRevE.101.010203

I. INTRODUCTION

Macroscopic description of numerous physical, chemical, and biological systems typically involves one or several partial differential equations (PDEs). In some instances, these PDEs represent a physical conservation law, and in others the PDEs are obtained by homogenization of an underlying microscopic description. The Navier-Stokes equation governing fluid flow and the diffusion equation governing heat or mass flux are examples that incorporate both approaches. Despite the differences in their origin, one thing remained constant for several centuries: PDE models were mainly derived from first principles. Their coefficients typically involve either fundamental physical constants, such as the Planck constant in the Schrödinger equation, or properties of the system, such as fluid viscosity or thermal conductivity, that can be either computed or measured independently.

An alternative approach (the data-driven discovery of mathematical models, where both the form of the model and the values of the coefficients are determined based solely on the available data) has emerged relatively recently [1–5]. In particular, sparse symbolic regression [6–8] has been applied successfully to identifying PDE models from data with minimal noise (i.e., a standard deviation of 1% or less of the data range). Unfortunately, since existing approaches based on sparse regression rely on the explicit evaluation of various candidate terms using local data, they all experience serious difficulties in the presence of higher noise levels characteristic of typical experimental measurements and generally fail to reconstruct PDE models involving higher-order derivatives.

Another limitation of existing approaches is that they require that all the variables present in the model be either directly observable or local functions of the directly observable data. For instance, using direct measurements of the fluid velocity $\mathbf{u}$, it is possible to reconstruct the vorticity equation [7], which involves $\mathbf{u}$ and the vorticity $\omega = \nabla \times \mathbf{u}$, but not the Navier-Stokes equation, which involves $\mathbf{u}$ and a latent variable: pressure. A recently introduced extension of the sparse regression method circumvents the latter limitation at the expense of raising the order of all of the derivatives [9].

An alternative approach that treats time evolution as a Gaussian process [10] was shown to be capable of reconstructing the coefficients in the two-dimensional (2D) Navier-Stokes equation without using the pressure field [11]. However, this approach assumes the model to be known *a priori* and exhibits noise sensitivity similar to that of sparse regression-based approaches. To the best of the authors' knowledge, no method currently exists that can robustly reconstruct PDEs involving latent variables (i.e., variables that cannot be measured) and/or derivatives of a high order using data with high levels of noise, which significantly limits the practical utility of data-driven approach to model discovery.

The present article removes a major roadblock for the data-driven approach in reconstructing PDE-based mathematical models by introducing a weak formulation of the sparse regression method, which addresses both of the limitations mentioned previously. Below we introduce our approach and illustrate it using three representative examples: the Kuramoto-Sivashinsky equation, a quasi-two-dimensional fluid flow, and the λ-ω reaction-diffusion system.

II. DATA-DRIVEN MODEL DISCOVERY

Models of continuous spatially distributed systems tend to have the form of a PDE

$$\sum_{n=0}^{N} c_n \mathbf{f}_n(\mathbf{u}, \partial_t \mathbf{u}, \partial_t^2 \mathbf{u}, \nabla \mathbf{u}, \nabla^2 \mathbf{u}, \dots) = 0, \tag{1}$$

where each of the terms depends on the system state $\mathbf{u}$ and its spatial and temporal derivatives of various orders and c_n are coefficients assumed to be constant in this study (an extension to coefficients depending on spatial and/or temporal coordinates is straightforward [6,8]). Symmetry and physical constraints can be used to narrow down the functional form of the terms that can appear in the model [9], and sparse regression can be used to discard unnecessary terms and determine a parsimonious form of the model and the values of the corresponding coefficients c_n.

We will illustrate the procedure using examples that involve a single term containing a temporal derivative of the state $\mathbf{u}$. The corresponding coefficient can be set to unity without loss of generality. Separating this term on the left-hand -side, we can rewrite (1) as

$$\partial_t^k \hat{D}\mathbf{u} = \sum_{n=1}^{N} c_n \mathbf{f}_n(\mathbf{u}, \nabla\mathbf{u}, \nabla^2\mathbf{u}, \dots), \tag{2}$$

where $\hat{D}$ is typically either an identity or a linear operator involving only spatial derivatives and k is the order of the temporal derivative. For instance, $k = 1$ and $\hat{D} = \mathbb{1}$ for the Navier-Stokes equation, $k = 2$ and $\hat{D} = \mathbb{1}$ for the wave equation, $k = 1$ and $\hat{D} = \nabla^2$ for the Orr-Sommerfeld equation, and so on.

To convert this to a linear algebra problem amenable to sparse regression, let us multiply the differential equation (2) by a weight $\mathbf{w}$ and integrate the result over a spatiotemporal domain Ω_k, then repeat this procedure for K different choices of Ω_k. This will yield a system

$$\mathbf{q}_0 = \sum_{n=1}^{N} c_n \mathbf{q}_n = Q\mathbf{c}, \tag{3}$$

where $Q = [\mathbf{q}_1 \dots \mathbf{q}_N]$ is the "library" and the "library terms" $\mathbf{q}_n \in \mathbb{R}^K$ are column vectors corresponding to different terms f_n in Eq. (2) with entries that correspond to a particular choice of $\mathbf{w}$ and Ω_k, e.g.,

$$q_n^k = \int_{\Omega_k} \mathbf{w} \cdot \mathbf{f}_n \, d\Omega. \tag{4}$$

The key advantage of this formulation compared to the local approach investigated previously [6–9] is that, by performing integration by parts, the action of derivatives can be transferred from the noisy data $\mathbf{u}$ onto the smooth weight $\mathbf{w}$, dramatically decreasing the effect of noise on terms involving high-order derivatives. Furthermore, the weight function can be chosen in such a way that the terms involving latent variables are eliminated, yielding a problem that can be solved using standard techniques.

A parsimonious model can finally be determined by choosing $K \geqslant N$ and using an iterative sparse regression algorithm such as SINDY [5]. Each iteration involves computing the solution

$$\tilde{\mathbf{c}} = Q^+ \mathbf{q}_0, \tag{5}$$

which minimizes the residual of the linear system defined by Eq. (3), where Q^+ denotes the pseudo-inverse of Q. This is followed by a thresholding procedure to remove dynamically irrelevant terms with $\|\tilde{c}_n \mathbf{q}_n\| < \gamma \|\mathbf{q}_0\|$ for sufficiently small γ (we choose $\gamma = 0.05$). To validate the results of regression, we use an ensemble of M cases with different random distributions of the K integration domains Ω_k relative to the spatiotemporal domain on which the data are available (we use $M = 30$ and $K = 100$).

Our approach is illustrated below using several examples that highlight different aspects of the problem. In the first two, we will assume that the form of the model is known, so that only the coefficients have to be determined. The last example illustrates how a parsimonious model can be identified via symbolic regression using a large library of candidate terms. In each case, we generate the synthetic data using the reference nonlinear PDE, add noise with standard deviation σ to this data, evaluate the integrals using the composite trapezoidal rule, and then solve the sparse regression problem to reconstruct the reference PDE. Note that, in all cases, the range of the data is $O(1)$, so that $\sigma = 1$ corresponds to 100% noise. Numerical codes used to generate the datasets and the MATLAB codes used to identify the governing equations using these datasets can be found in the GitHub repository [12].

III. HIGH-ORDER DERIVATIVES

The Kuramoto-Sivashinsky equation

$$\partial_t u + u\partial_x u + \partial_x^2 u + \partial_x^4 u = 0, \tag{6}$$

describes the chaotic dynamics of laminar flame fronts [13], reaction-diffusion systems [14], and coating flows [15]. This is a notable example of a nonlinear PDE that involves high-order partial derivatives, which has made it difficult to accurately reconstruct from noisy data. Rearranging this PDE into the form of Eq. (2), we find $c_1 = c_2 = c_3 = -1$.

Since the Kuramoto-Sivashinsky equation involves a scalar variable u, it can be converted to weak form by integrating its product with a scalar weight w over a set of different integration domains

$$\Omega_k = \{(x, t) : |x - x_k| \leqslant H_x, |t - t_k| \leqslant H_t\} \tag{7}$$

centered around randomly chosen points (x_k, t_k). This yields the system (3) with library terms whose elements are given by

$$q_0^k = \int_{\Omega_k} w\partial_t u \, d\Omega, \qquad q_1^k = \int_{\Omega_k} wu\partial_x u \, d\Omega,$$
$$q_2^k = \int_{\Omega_k} w\partial_x^2 u \, d\Omega, \qquad q_3^k = \int_{\Omega_k} w\partial_x^4 u \, d\Omega. \tag{8}$$

Integration by parts can be used to move all derivatives from the noisy field u onto a smooth noiseless w, yielding

$$q_0^k = -\int_{\Omega_k} u\partial_t w \, d\Omega, \qquad q_1^k = -\int_{\Omega_k} \frac{1}{2} u^2 \partial_x A \, d\Omega,$$
$$q_2^k = \int_{\Omega_k} u\partial_x^2 w \, d\Omega, \qquad q_3^k = \int_{\Omega_k} u\partial_x^4 w \, d\Omega, \tag{9}$$

provided w satisfies the conditions required for the boundary terms to vanish. Specifically, w (and its derivatives up to third order in space) should vanish along the boundary $\partial\Omega_k$. To satisfy these boundary conditions, we chose

$$w = (\underline{x}^2 - 1)^p (\underline{t}^2 - 1)^q, \tag{10}$$

where $p \geqslant 4$, $q \geqslant 1$ are integers and the underbar denotes nondimensionalized variables $\underline{x} = (x - x_k)/H_x$ and $\underline{t} = (t - t_k)/H_t$. Of course, many other choices for w are possible as well.

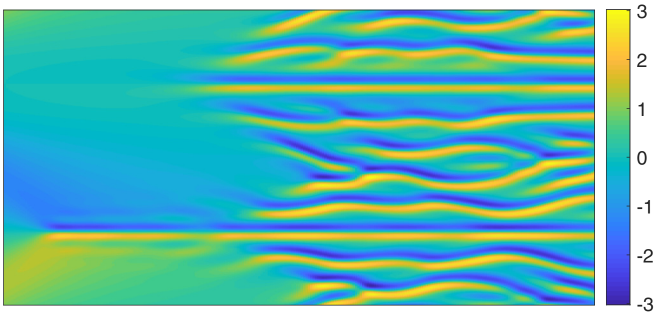

FIG. 1. Space-time plot of the solution to the Kuramoto-Sivashinsky equation. The x axis is vertical and the t axis is horizontal. The figures below use a similar colormap.

The linear system (3) can now be constructed by evaluating the integrals in Eq. (9) over a set of domains Ω_k. To test our sparse regression approach, we generated synthetic data by solving the Kuramoto-Sivashinsky equation numerically. To enable direct comparison to the results of Rudy *et al.* [7], we used the same integrator [16] to compute the solution of Eq. (6) on a spatiotemporal domain of size $L_x = 32\pi$ and $L_t = 100$ using a grid with the same density $\Delta x = 0.0982$ and $\Delta t = 0.4$; the solution is shown in Fig. 1. Gaussian noise with standard deviation σ was then added to u at each grid point, after which the integrals in Eq. (9) were evaluated over integration domains with dimensions $H_x \approx 24.5$, $H_t = 20$. The weight function used the exponents $p = 4$ and $q = 3$.

The results for different noise levels are shown in Fig. 2, with the accuracy of the model reconstruction quantified by the relative errors

$$\Delta c_n = \left|\frac{\tilde{c}_n - c_n}{c_n}\right|, \tag{11}$$

where c_n are the coefficients used to generate the numerical data and $\tilde{c}_n$ are the coefficients estimated from noisy data by our sparse regression algorithm. Here and below, the symbols and the error bars show the mean values and the full range of the results, respectively, for the entire ensemble. Note that the reconstruction remains essentially unaffected by noise, with an error of about 1% or below, until the noise level exceeds 10%. This is a dramatic improvement compared to the original study [7], which yielded errors of over 50% for all of the coefficients with just 1% noise.

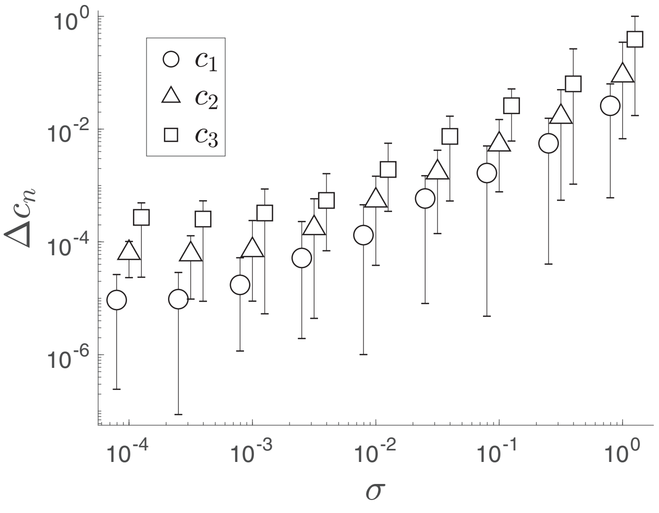

FIG. 2. The accuracy of parameter reconstruction for the Kuramoto-Sivashinsky equation as a function of the noise amplitude.

IV. LATENT VARIABLES

To illustrate our approach applied to systems with latent variables, we next consider a flow in a thin layer of fluid driven by a steady but spatially nonuniform force $\mathbf{f}$. The flow can be described using a generalization of the 2D Navier-Stokes equation

$$\partial_t \mathbf{u} = c_1(\mathbf{u}\cdot\nabla)\mathbf{u} + c_2\nabla^2\mathbf{u} + c_3\mathbf{u} - \nabla p + \mathbf{f}, \tag{12}$$

where $\mathbf{u} = \hat{x}u_x + \hat{y}u_y$ is the flow field, which is considered to be incompressible, p is the pressure, and the constants c_1, c_2, and c_3 describe, respectively, the depth-averaged effects of inertia and viscosity in the horizontal and vertical direction [17,18]. In this example, both p and $\mathbf{f}$ are assumed to be latent variables that cannot be measured.

To convert Eq. (12) to weak form, we multiply it by a vector field $\mathbf{w}$ and integrate the result by parts over a (now three-dimensional) spatiotemporal domain Ω_k of size $2H_x \times 2H_y \times 2H_t$. Assuming again that the boundary terms vanish, for the linear terms we immediately find

$$\begin{aligned} q_0^k &= \int_{\Omega_k} \mathbf{w}\cdot\partial_t\mathbf{u}\,d\Omega = -\int_{\Omega_k}\mathbf{u}\cdot\partial_t\mathbf{w}\,d\Omega,\\ q_2^k &= \int_{\Omega_k}\mathbf{w}\cdot\nabla^2\mathbf{u}\,d\Omega = \int_{\Omega_k}\mathbf{u}\cdot\nabla^2\mathbf{w}\,d\Omega,\\ q_3^k &= \int_{\Omega_k}\mathbf{w}\cdot\mathbf{u}\,d\Omega. \end{aligned} \tag{13}$$

The nonlinear term can be rewritten in a similar way using the incompressibility condition $\partial_i u_i = 0$ (where summation over repeated indices is implied):

$$\begin{aligned} q_1^k &= \int_{\Omega_k} w_i u_j \partial_j u_i\,d\Omega = -\int_{\Omega_k} u_i\partial_j(w_i u_j)\,d\Omega\\ &= -\int_{\Omega_k} u_i u_j \partial_j w_i\,d\Omega = -\int_{\Omega_k}\mathbf{u}\cdot(\mathbf{u}\cdot\nabla)\mathbf{w}\,d\Omega. \end{aligned} \tag{14}$$

Finally, for the terms involving the latent variables, we find

$$\begin{aligned} q_4^k &= \int_{\Omega_k}\mathbf{w}\cdot\nabla p\,d\Omega = -\int_{\Omega_k} p\nabla\cdot\mathbf{w}\,d\Omega,\\ q_5^k &= \int_{\Omega_k}\mathbf{w}\cdot\mathbf{f}\,d\Omega. \end{aligned} \tag{15}$$

In order for the boundary terms to vanish on a rectangular domain Ω_k centered at (x_k, y_k, t_k), we need to have $\mathbf{w} = 0$ on $\partial\Omega$, as well as $\partial_x\mathbf{w} = 0$ at $\underline{x} = \pm 1$ and $\partial_y\mathbf{w} = 0$ at $\underline{y} = \pm 1$, where the underbar denotes rescaled variables $\underline{x} = (x - x_k)/H_x$, $\underline{y} = (y - y_k)/H_y$, and $\underline{t} = (t - t_k)/H_t$. Next the dependence on the pressure field and the steady forcing can be eliminated by additionally requiring that

$$\nabla\cdot\mathbf{w} = 0 \tag{16}$$

and

$$\int_{-1}^{1} \mathbf{w}\, d\underline{t} = 0. \tag{17}$$

All of the above conditions on $\mathbf{w}$ can be satisfied by setting $\mathbf{w} = \nabla \times (\psi \hat{z}) = \hat{x}\partial_y \psi - \hat{y}\partial_x \psi$, where

$$\psi = \sin(\pi \underline{t})(\underline{x}^2 - 1)^p(\underline{y}^2 - 1)^p \tag{18}$$

and $p \geqslant 3$ (we used $p = 3$ in this study). This yields $q_4^k = q_5^k = 0$ and

$$\begin{aligned}
q_0^k &= -\int_{\Omega_k} (u_x\partial_y - u_y\partial_x)\partial_t \psi \, d\Omega, \\
q_1^k &= \int_{\Omega_k} \left[u_x u_y\left(\partial_y^2 - \partial_x^2\right) + \left(u_x^2 - u_y^2\right)\partial_{xy}\right]\psi \, d\Omega, \\
q_2^k &= \int_{\Omega_k} (u_x\partial_y - u_y\partial_x)\nabla^2\psi \, d\Omega, \\
q_3^k &= \int_{\Omega_k} (u_x\partial_y - u_y\partial_x)\psi \, d\Omega.
\end{aligned} \tag{19}$$

As in the case of the Kuramoto-Sivashinsky equation, the linear system (3) can now be constructed by evaluating the integrals in Eq. (19) over a set of domains Ω_k. Note that this linear system involves neither the derivatives of the noisy observable data (components of the $\mathbf{u}$ field) nor the latent variables (p and $\mathbf{f}$ fields).

To test our approach, we generated synthetic data $\mathbf{u}$ by solving Eq. (12) with the parameters $c_1 = -0.826$, $c_2 = 0.0487$, and $c_3 = -0.157$, which correspond to the experimental setup of Kolmogorov-like flow described in Ref. [18]. In the experiment, the forcing field $\mathbf{f} = f(x, y)\hat{x}$ is produced by an array of long bar magnets with alternating polarity and width equal to unity in nondimensional units; hence $f(x, y)$ is approximately uniform in the x direction and nearly periodic in the y direction [cf. Fig. 3(a)], with the "period" equal to 2 units. Forcing with amplitude $\max_{x,y} |f(x, y)| = 1.0649$ generates a weakly turbulent flow [a representative snapshot is shown in Fig. 3(b)], which was computed using the numerical integrator described in Ref. [18] on a domain of size $L_x = 14$, $L_y = 18$, $L_t \approx 920$, and a computational grid with $\Delta x_c = \Delta y_c = 0.025$ and $\Delta t_c \approx 0.02$. The data were then subsampled on a coarser grid with spacing $\Delta x = \Delta y = 0.1$ and $\Delta t = 0.2302$, and Gaussian random noise with variance σ was added to both components of $\mathbf{u}$. The integrals in Eq. (19) were evaluated over domains Ω_k of size $H_x = 11.2$, $H_y = 14.4$, and $H_t \approx 34.5$.

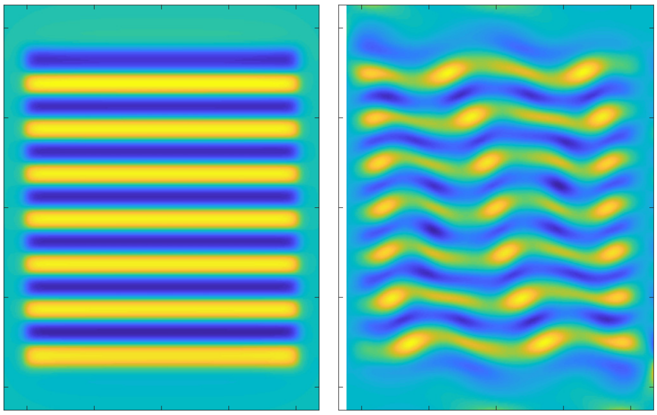
(a) (b)

FIG. 3. The Kolmogorov-like flow: (a) the forcing profile f and (b) snapshot of the vorticity $\omega = \partial_x u_y - \partial_y u_x$. The x axis is horizontal and the y axis is vertical.

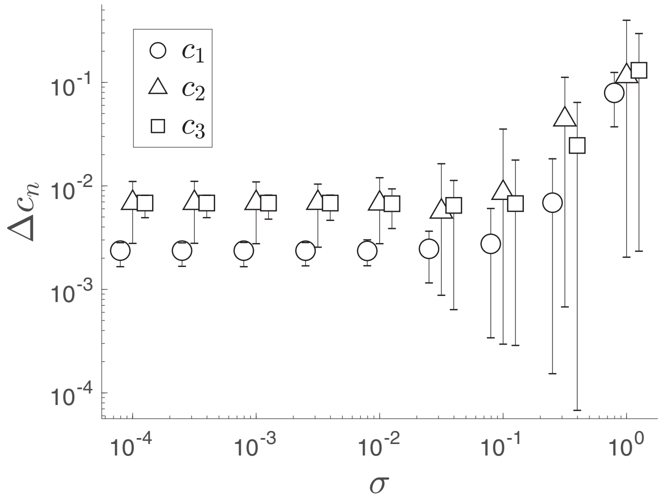

FIG. 4. The accuracy of parameter reconstruction for the 2D Kolmogorov-like flow model as a function of the noise amplitude.

As Fig. 4 illustrates, our approach successfully reconstructs the reference PDE (12). Just like in the case of the Kuramoto-Sivashinsky equation, noise up to 10% does not meaningfully affect the accuracy of model reconstruction, with the coefficients c_1, c_2, and c_3 estimated to within 1% or better. In fact, even with 100% noise, the coefficients can still be estimated to within roughly 10%. For reference, the experimental data [18] obtained using particle image velocimetry has roughly 3% noise, at which level local sparse regression [9] failed completely.

V. SPARSE REGRESSION

Finally, as an example of how the proposed approach could be used in the context of sparse regression, we consider the λ-ω reaction-diffusion system [19]

$$\begin{aligned}
\partial_t u &= D\nabla^2 u + \lambda u - \omega v, \\
\partial_t v &= D\nabla^2 v + \omega u + \lambda v,
\end{aligned} \tag{20}$$

where $\omega = -\beta(u^2 + v^2)$, $\lambda = 1 - u^2 - v^2$, and $\beta = 1$ and $D = 0.1$ are constants. This system can be cast in the form of Eq. (2) by defining a vector $\mathbf{u} = [u, v]$. To test our approach, we applied sparse regression to a generalization of Eq. (20), where the reaction terms are given by polynomials in u and v up to third order. In total, the generalized model involves a total of 20 different terms (two diffusion terms and 18 polynomial terms). Correspondingly, 20 unknown coefficients need to be determined.

The sparse regression problem for the λ-ω system can be block-diagonalized by using a weight function $\mathbf{w} = [w, 0]$ (or $\mathbf{w} = [0, w]$) to reconstruct the first (or second) equation

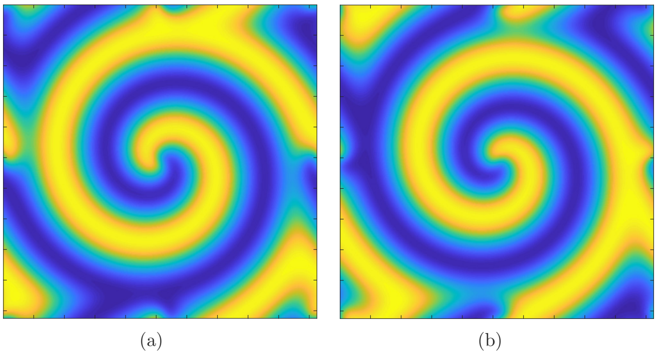

FIG. 5. A typical snapshot of the fields (a) u and (b) v for the λ-ω reaction diffusion system. The x axis is horizontal and the y axis is vertical.

in Eq. (20), yielding two independent linear systems (3) with ten library terms each. The integration domains Ω_k are three dimensional as in the previous example. The integrals involving terms such as $u^\alpha v^\beta$ do not require integration by parts. The two integrals involving the Laplacian terms are integrated by parts twice to get rid of derivatives on u and v, e.g.,

$$q_1^k = \int_{\Omega_k} w\nabla^2 u \, d\Omega = \int_{\Omega_k} u\nabla^2 w \, d\Omega. \tag{21}$$

In both cases, the corresponding boundary terms vanish if we choose

$$w = (\underline{x}^2 - 1)^p(\underline{y}^2 - 1)^p(\underline{t}^2 - 1)^q, \tag{22}$$

where $p \geqslant 2$ and $q \geqslant 1$ (we chose $p = 2$ and $q = 1$).

The synthetic data were obtained by computing the solution of Eq. (20) using the integrator employed in Ref. [7]; a typical snapshot is shown in Fig. 5. The computational domain of size $L_x = 20$, $L_y = 20$, $L_t = 10$ was discretized using a grid with spacing $\Delta x = \Delta y = 0.0391$ and $\Delta t = 0.05$, and Gaussian random noise with standard deviation σ was added to both u and v at each grid point. The dimensions of the integration domains Ω_k were chosen as $H_x = H_y \approx 1$ and $H_t = 1.25$.

The results of sparse regression are shown in Fig. 6. We find that, for noise levels of up to 5%, the model was reconstructed correctly (with no spurious or missing terms) for each distribution of Ω_k in our ensemble, with all parameters estimated to an accuracy of better than 1%. With 10% noise, the model is identified correctly in about 95% of cases, and at 30% noise, the model is identified correctly in about 20% of cases, with the remaining cases featuring spurious terms (linear in u and v) that are not present in the λ-ω model. For reference, sparse regression based on local evaluation of derivatives [7] failed to correctly identify this model, generating spurious terms in the presence of as little as 1% noise.

It should be noted that using ensemble sparse regression makes it easy to detect the presence of spurious (missing) terms and eliminate (add) them while still preserving the accuracy with which all of the correct terms are estimated (in our case, about 3% for the worst case offenders with 10% noise). It is also worth pointing out that, unlike the standard approach [7], weak formulation requires no intermediate noise reduction.

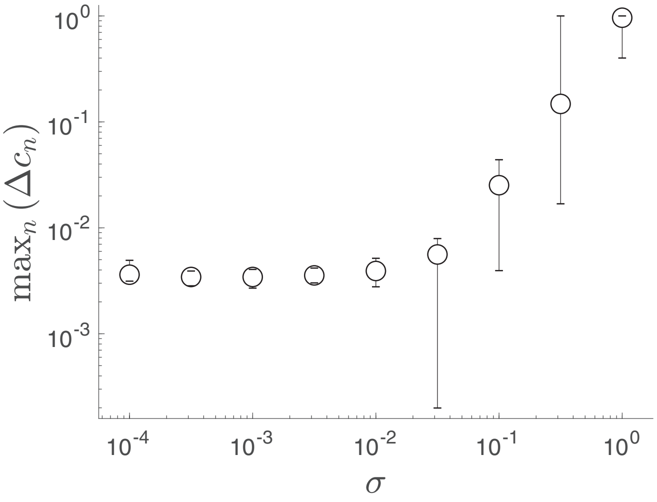

FIG. 6. The accuracy of parameter reconstruction for the λ-ω reaction-diffusion system as a function of the noise amplitude. Shown is the largest error, which corresponds to one of the diffusion coefficients.

Predictions of all three reconstructed models are compared to synthetic data in the Supplemental Material [20].

VI. DISCUSSION

The examples presented here illustrate the power of the weak formulation of sparse regression applied to noisy and/or incomplete data. For instance, high-order PDEs such as the Kuramoto-Sivashinsky equation simply cannot be reconstructed with meaningful accuracy from data with realistic levels of noise using the original (differential) form of the model. The main culprit is the term in the model involving a fourth-order derivative, which is extremely sensitive even to minute amounts of noise. The weak formulation involves integrals of the data rather than derivatives, which makes it much more robust with respect to noise. While the weak formulation may not eliminate *all* of the derivatives in some models (e.g., in nonlinear terms), it can reduce the order of the derivatives that remain, which is extremely beneficial when noisy data are involved.

We also demonstrated that the weak formulation of sparse regression can be applied successfully to models with latent variables, as in the example of the fluid flow in a thin layer, where neither the pressure field nor the forcing field are accessible. Needless to say, the weak formulation by itself simply eliminates rather than reconstructs the terms that involve the latent variables. One needs to impose additional physical constraints [9] to determine their functional form. Nonetheless, the approach presented here has substantial advantages compared to the method described in Ref. [9], which involves taking additional spatial and/or temporal derivatives of the model equation to eliminate the latent variables. As discussed previously, the higher the order of the derivatives, the more sensitive the sparse regression is to noise. As a result, the model (12) could only be reconstructed with acceptable accuracy in that study for noise levels of 0.01% or less. The present approach gives better accuracy for data with as much as 30% noise.

In conclusion, let us point out that we made no attempt to optimize our approach here. Several options are available to make it even more robust and accurate [21]. As an example, the size of the integration domains Ω_k could be varied relative to the size of the spatiotemporal domain on which the data are available. Furthermore, we only used a single weight function, while in principle one could also use a set of different weight functions $\mathbf{w}_j$. Additionally, the shape of the weight functions could be optimized to improve the accuracy even compared to the substantial results presented here. For instance, simply increasing the powers p and q beyond the minimal possible values (determined, respectively, by the highest order of the spatial and temporal derivatives in the model) can reduce the error in estimating the coefficients of the model by orders of magnitude. In contrast, we found the details of the sparse regression procedure itself to have a relatively minor impact on the results.

ACKNOWLEDGMENTS

This material is based upon work supported by NSF under Grant No. CMMI-1725587 and ARO under Grant No. W911NF-15-10471. D.R.G. gratefully acknowledges the support of the Letson Undergraduate Research Scholarship.

[1] J. P. Crutchfield and B. S. McNamara, Equation of motion from a data series, Complex Syst. **1**, 417 (1987).

[2] J. Bongard and H. Lipson, Automated reverse engineering of nonlinear dynamical systems, Proc. Natl. Acad. Sci. USA **104**, 9943 (2007).

[3] C. Yao and E. M. Bollt, Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems, Physica D: Nonlinear Phenomena **227**, 78 (2007).

[4] I.-C. Chou and E. O. Voit, Recent developments in parameter estimation and structure identification of biochemical and genomic systems, Math. Biosci. **219**, 57 (2009).

[5] 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**, 3932 (2016).

[6] D. Xu and O. Khanmohamadi, Spatiotemporal system reconstruction using fourier spectral operators and structure selection techniques, Chaos **18**, 043122 (2008).

[7] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, Data-driven discovery of partial differential equations, Science Advances **3**, e1602614 (2017).

[8] X. Li, L. Li, Z. Yue, X. Tang, H. U. Voss, J. Kurths, and Y. Yuan, Sparse learning of partial differential equations with structured dictionary matrix, Chaos **29**, 043130 (2019).

[9] P. A. K. Reinbold and R. O. Grigoriev, Data-driven discovery of partial differential equation models with latent variables, Phys. Rev. E **100**, 022219 (2019).

[10] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Numerical gaussian processes for time-dependent and nonlinear partial differential equations, SIAM J. Sci. Comput. **40**, A172 (2018).

[11] M. Raissi and G. E. Karniadakis, Hidden physics models: Machine learning of nonlinear partial differential equations, J. Comput. Phys. **357**, 125 (2018).

[12] https://github.com/pakreinbold/PDE_Discovery_Weak_Formulation

[13] G. Sivashinsky, Nonlinear analysis of hydrodynamic instability in laminar flames: I. derivation of basic equations, Acta Astronaut. **4**, 1177 (1977).

[14] Y. Kuramoto and T. Tsuzuki, Persistent propagation of concentration waves in dissipative media far from thermal equilibrium, Prog. Theor. Phys. **55**, 356 (1976).

[15] G. I. Sivashinsky and D. Michelson, On irregular wavy flow of a liquid film down a vertical plane, Prog. Theor. Phys. **63**, 2112 (1980).

[16] A.-K. Kassam and L. N. Trefethen, Fourth-order time-stepping for stiff pdes, SIAM J. Sci. Comput. **26**, 1214 (2005).

[17] B. Suri, J. Tithof, R. Mitchell, R. O. Grigoriev, and M. F. Schatz, Velocity profile in a two-layer Kolmogorov-like flow, Phys. Fluids **26**, 053601 (2014).

[18] J. Tithof, B. Suri, R. K. Pallantla, R. O. Grigoriev, and M. F. Schatz, Bifurcations in a quasi-two-dimensional Kolmogorov-like flow, J. Fluid Mech. **828**, 837 (2017).

[19] N. Kopell and L. N. Howard, Plane wave solutions to reaction-diffusion equations, Stud. Appl. Math. **52**, 291 (1973).

[20] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.101.010203 for a comparison of the three reconstructed models and the synthetic data.

[21] D. R. Gurevich, P. A. K. Reinbold, and R. O. Grigoriev, Robust and optimal sparse regression for nonlinear PDE models, Chaos **29**, 103113 (2019).