Robust and optimal sparse regression for nonlinear PDE models

Daniel R. Gurevich,1 Patrick A. K. Reinbold,1 and Roman O. Grigoriev1

School of Physics, Georgia Institute of Technology, Atlanta, GA 30332, USA

(Dated: 21 July 2019)

This paper investigates how models of spatiotemporal dynamics in the form of nonlinear partial differential equations can be identified directly from noisy data using a combination of sparse regression and weak formulation. Using the 4th-order Kuramoto-Sivashinsky equation for illustration, we show how this approach can be optimized in the limits of low and high noise, achieving accuracy that is orders of magnitude better than what existing techniques allow. In particular, we derive the scaling relation between the accuracy of the model, the parameters of the weak formulation, and the properties of the data, such as its spatial and temporal resolution and the level of noise.

Keywords: data-driven discovery, machine learning, sparse regression, partial differential equations

In recent years, data-driven discovery of mathematical models of spatially extended systems described by nonlinear PDEs has emerged as a promising alternative to more traditional modeling approaches. Existing approaches to model discovery such as sparse regression have several major weaknesses, however. Most notably, they break down for data with high levels of noise and have to be tuned empirically to produce meaningful results, making them ill-suited for analyzing experimental data. We show how these weaknesses can be addressed using a weak formulation of the model PDE. The weak formulation has substantial freedom that makes it extremely powerful and flexible, but the question arises of how this freedom can be used to robustly obtain the most accurate model. This question is addressed here for the first time.

I. INTRODUCTION

Partial differential equations (PDEs) provide a natural description for the temporal evolution of spatially extended systems in various fields of science and engineering. Historically and practically important examples include wave equations arising in many areas of physics, the Schrödinger equation in quantum mechanics, the Navier-Stokes equations in fluid dynamics, and reaction-diffusion equations used to model physical, chemical, or biological systems. In the past, models of such systems were almost always constructed from first principles or using a suitable empirical approach. However, in recent years, a data-driven paradigm for learning the dynamics has emerged, which leverages the modern prevalence of data and computational power to create models when the underlying governing laws have eluded first-principles derivation.

Many indirect methods for learning the dynamics that do not require a PDE have been proposed. Notable examples include equation-free modeling,14–16 artificial neural networks,17–19 dynamic mode decomposition,20 and Koopman operator approaches,21 balanced truncation,22 and resolvent-based analysis.23 While these techniques can provide an economical approximate description of the dynamics, this is done at the cost of losing the mathematical structure that affords physical intuition or interpretability. Symbolic regression, which was originally used to derive nonlinear ordinary differential equations describing low-dimensional systems,24 offers an enticing alternative by allowing construction of exact models and discovery of conservation laws. The genetic algorithms used in these earlier studies are however computationally expensive, preventing application of this approach to high-dimensional systems. Thus, the recent emergence of a sparse regression approach for model discovery25–27 has made a significant impact. Applied to spatially extended systems, this approach allows data-driven discovery of governing equations in the form of PDEs by evaluating a library of candidate terms containing partial derivatives at a large number of points and using a regularized regression procedure to compute the coefficients of each term and select a parsimonious model.

Sparse regression has proven computationally efficient and capable of reconstructing numerous canonical PDE,25–27 but it faces serious difficulties when used for analysis of experimental data. One complication is that the proper choice of parsimonious model is often unclear. In many implementations, it relies on a manual Pareto analysis to balance model accuracy and complexity or on an automatic but complex thresholding procedure (e.g. sequential threshold ridge regression) that tends to be sensitive to the choice of parameters. More importantly, existing sparse regression methods often suffer from low accuracy even in the absence of noise and completely break down at noise levels characteristic of realistic applications. This is because they inherently require explicit numerical evaluation of partial derivatives of the data, which is a notably ill-conditioned problem.

In this paper, we present a weak formulation of the sparse regression problem that eliminates this fundamental issue. We also suggest a simple thresholding procedure that can always identify the correct form of the governing PDE even in the presence of extremely high noise. Finally, we explore how this extremely flexible and robust
approach can be optimized and tuned to the properties of the underlying data set to maximize accuracy. This paper has the following structure. Section II describes our approach and the system used to test it. Results are presented and interpreted in Section III, and conclusions are discussed in Section IV.

II. METHODS

We consider the problem of using the data \( u(x, t) \) to identify a parsimonious mathematical model in the form of a PDE

\[
\sum_{n=1}^{N} c_n f_n(x, t, u, \partial_t u, \partial_t^2 u, \nabla u, \nabla^2 u, \cdots) = 0 \quad (1)
\]

where each term in the sum is a function of \( u \) and its partial derivatives in space and time with constant coefficients \( c_n \). In most applications, the form of the basis functions \( f_n \) can be restricted based on physical considerations, such as symmetries, conservation laws, etc. Typically, \( f_n \) are taken to be products of powers of independent variables \( (x, t) \) and dependent variables \( (u \) and its various derivatives), although the form can be arbitrary in theory. Our goal is to determine the constants \( c_n \) for the terms that should be present in the model while eliminating the dynamically insignificant and thus likely spurious terms. Sparse regression aims to convert the PDE (1) to a tractable (and ideally, robust) linear algebra problem. Conventionally this is done by evaluating all of the terms in the PDE at a random collection of points \( (x_k, t_k) \) using finite differences, spectral methods, or polynomial approximation. All of these approaches are extremely sensitive to noisy, especially when high-order derivatives are present. We will instead pursue a weak formulation of the problem that can be obtained by multiplying (1) by a weight \( w_i(x, t) \) and then integrating the result over a domain \( \Omega_k \). Repeating the process for \( K \) distinct combinations of weight functions and integration domains yields the linear system

\[
Qc = 0 \quad (2)
\]

where \( c = [c_1, \ldots, c_N]^T \) and \( Q = [q_1, \ldots, q_N] \) is a “library” matrix, with each column \( q_n \) consisting of the integrals of the function \( f_n \) with all \( K \) combinations of weights \( w_i \) and domains \( \Omega_k \).

Note that there is an extra degree of freedom in (2) corresponding to the normalization of \( c \). Conventionally this is dealt with by assuming that \( f_0 = \partial_t u \), setting \( c_1 = 1 \), and solving the overdetermined system that corresponds to the choice \( K \gg N \) using least squares or some regularized version of it. This is however not always a valid assumption: it is usually unknown \textit{a priori} whether any given temporal derivative should be included in the PDE at all, whereas in this case a particular term is forced into the model. Moreover, even if this term should be present in the model, the regression effectively assumes that the time derivative was computed without error, which reduces the practical accuracy of the procedure.

We will therefore not make the assumption that the model has the form of an evolution equation and consider the linear problem (2) in its most general form. The normalization of \( c \) can be fixed by adding an extra row with arbitrary nonzero elements to \( Q \), after which the resulting equation (2) can be solved by ordinary least squares. A more elegant solution pursued in the present study is to instead compute \( c \) as the right singular vector of \( Q \) corresponding to the smallest singular value. Note that this corresponds to the solution of a constrained least squares problem for \( Q^T Q c = 0 \):

\[
c = \arg \min_{\|c\|_1=1} \|Q^T Q c\|_1.
\]

Once a suitable solution has been obtained by further constraining the problem, the resulting parsimonious model can be rewritten in the form of an evolution equation by solving for a term such as \( \partial_t u \) (or \( \partial_t^2 u \) for a wave equation).

To obtain a parsimonious model, we employ an iterative procedure to eliminate unnecessary terms from (1). At each step \( i \), singular value decomposition is used to obtain the solution \( c^i \) given the matrix \( Q^i \), and the residual \( \eta^i = \|Q^i c^i\|_1 \) is computed. We then find the term with the smallest \( \|c_n^i q_n^i / \|q_n^i\|_2 \) and construct \( Q^{i+1} \) by eliminating the column \( q_n^i \) from \( Q^i \). The corresponding term is eliminated from the model if \( \eta^{i+1} < \gamma \eta^i \), where \( \gamma > 1 \) is some fixed constant (we use \( \gamma = 1.4 \) in the present study). The iteration terminates at step \( i \) if \( \eta^{i+1} > \gamma \eta^i \), yielding a parsimonious model. We find that this method compares favorably to alternatives such as sequential threshold ridge regression, as it robustly eliminates spurious terms without requiring extremely careful choice of parameters. Moreover, the sparsification parameter has a simple interpretation: \( \gamma - 1 \) is the maximum acceptable relative increase in the residual resulting from discarding a single library term.

We illustrate the advantages of our approach by applying it to the Kuramoto-Sivashinsky equation:

\[
c_1 \partial_t u + c_2 u \partial_x u + c_3 \partial_x^2 u + c_4 \partial_x^4 u = 0 \quad (4)
\]

which has posed a significant challenge in past studies of sparse regression because it contains a fourth-order partial derivative that is difficult to evaluate numerically with adequate accuracy. Here \( c_1 = \cdots = c_4 = 1 \) are all constants, although our approach can easily be extended even to the case when these coefficients are functions of time and/or space, as discussed below. Since this is a scalar equation in one spatial and one temporal dimension, we use scalar weight functions \( w^j(x, t) \). If we denote
the terms in the model \[\{\] by \(f_1, \ldots, f_4\), then

\[
q_1^{jk} = \int_{\Omega_k} w_j \partial_t w_d \Omega, \quad q_2^{jk} = \int_{\Omega_k} w_j \partial_x u d\Omega, \\
q_3^{jk} = \int_{\Omega_k} w_j \partial_x^2 w_d \Omega, \quad q_4^{jk} = \int_{\Omega_k} w_j \partial_x^4 u d\Omega, 
\]  
(5)

where \(d\Omega = dx \, dt\). The key feature of the weak formulation is that it can almost always be used to completely eliminate, or at least reduce the order of, the derivatives acting on the noisy data by integrating by parts. In our particular case,

\[
q_1^{jk} = -\int_{\Omega_k} u \partial_t w_j \, d\Omega, \quad q_2^{jk} = -\frac{1}{2} \int_{\Omega_k} u^2 \partial_x w_j \, d\Omega, \\
q_3^{jk} = \int_{\Omega_k} u \partial_x^2 w_j \, d\Omega, \quad q_4^{jk} = \int_{\Omega_k} u \partial_x^4 w_j \, d\Omega 
\]

(6)

under the assumption that \(w_j\) and its first three partial derivatives with respect to \(x\) vanish on the boundary \(\partial\Omega_k\). In our implementation, we use the composite trapezoidal rule to evaluate the integrals numerically.

Note that although this particular PDE features constant coefficients, terms with variable coefficients can be treated in a similar manner. For instance, suppose that the coefficient of the term \(\partial_x^2 u\) is a function of \(x\) and \(t\) that can be expanded in some (finite) basis as

\[
c_4(x, t) = \sum_p c'_p g_p(x, t) 
\]

(7)

with some constants \(c'_p\). Then

\[
\int_{\Omega_k} w_j c_4 \partial_x^2 u \, d\Omega = \sum_p c'_p q_p^{jk}, 
\]

(8)

where

\[
q_p^{jk} = \int_{\Omega_k} u \partial_x^2 (g_p w_j) \, d\Omega. 
\]

(9)

Sparse regression for a model including such a term would then simply require expanding the library \(Q\) to include additional columns \(q_p\) with entries \(q_p^{jk}\). In this case as well, no derivatives of the noisy \(u\) are used in finding the elements of \(Q\).

Although in principle integration domains of any shape can be used, here we will only consider rectangular domains of a fixed size

\[
\Omega_k = \{(x, t) : |x - x_k| \leq H_x, |t - t_k| \leq H_t\} 
\]

(10)

where the centers \((x_k, t_k)\) of the rectangles \(\Omega_k\) are chosen randomly. Similarly, there are many possible choices for the weight functions satisfying the boundary conditions on \(\partial\Omega_k\); we focus on functions of the form

\[
w_j = (\alpha^2 - 1)^\beta e^{\pm i\pi x} e^{\pm i\pi t}, 
\]

(11)

where \(x = (x - x_k)/H_x, t = (t - t_k)/H_t\) are nondimensionalized independent variables and \(\alpha \geq 4, \beta \geq 1, l \geq 0,\) and \(m \geq 0\) are integers. Note that there are four weight functions (corresponding to the four different choices of the signs in the exponentials) for each pair of nonzero \(l\) and \(m\). The integrals \(q_p^{jk}\) are all of the form

\[
F_n^m = \int_{-\frac{l}{2}}^{\frac{l}{2}} dt \int_{-\frac{m}{2}}^{\frac{m}{2}} dz f_n^m(x, t) e^{\pm il\pi x} e^{\pm im\pi t}, 
\]

(12)

where

\[
f_n^m(x, t) = f_n(u, x, t)(\alpha^2 - 1)^\alpha (l^2 - 1)^\beta, 
\]

(13)

so \(F_n^m\) are the coefficients of the two-dimensional Fourier series for \(f_n^m(x, t)\). Although \(f_n(u, x, t)\) is not periodic on \(\Omega_k\), the functions \(f_n^m(x, t)\) are. Moreover, \(f_n^m(x, t)\) has at least \(\alpha - 1\) continuous derivatives in \(x\) and \(\beta - 1\) continuous derivatives in \(t\), so the Fourier coefficients decay according to \(F_n^m \sim l^{-\alpha} m^{-\beta}\). The powers \(\alpha\) and \(\beta\) therefore control the width of the Fourier spectrum of the entries \(q_p^{jk}\) in the library \(Q\), while the choice of \(l\) and \(m\) allows us to tune the frequencies of the weights to the spectral properties of the data. The convergence rate of Fourier series turns out to control the accuracy with which the integrals are evaluated using data that are available only on a discrete grid. For simplicity, we will assume that the same weight functions are integrated on every domain. It is possible to use either weight functions involving only a single pair of frequencies (e.g., \(l\) and \(m\)) or a range of frequencies in space and/or time.

To test our sparse regression approach, we computed a solution of the Kuramoto-Sivashinsky equation, using the integrator described in Ref. 14 to generate data on a physical domain with dimensions \(L_x = 32\pi\) and \(L_t = 500\). The numerical integration generated data with spatial resolution \(\Delta x = 0.0491\) using a time step \(\Delta t = 0.005\), which was then downsampled to a lower spatial resolution \(\delta_x\) and temporal resolution \(\delta_t\). Unless noted otherwise, the results presented below are for \(\delta_x = 0.1964\) and \(\delta_t = 1\). For reference, the solution has a correlation length \(\ell_x \sim 1.67 \approx 8.5\delta_x\) and correlation time \(\ell_t \approx 8 \approx 8\delta_t\). To test the effects of noise, Gaussian noise with standard deviation \(\sigma_u\) was added to the data for various choices of \(\sigma\), where \(s_u \approx 1.3\) is the sample standard deviation of \(u\) on the whole domain.

To test the ability of the algorithm to eliminate spurious terms, in addition to the terms present in the Kuramoto-Sivashinsky equation \(\{\] we also included terms \(\partial_x u, \partial_x^2 u, u, u^2, t^2, 1\) (which represents a hypothetical forcing) in our library. The corresponding integrals were rewritten using integration by parts to remove derivatives acting on \(u\), as described previously. In the next section, we quantify the performance of our sparse regression approach using two key metrics: how well the algorithm can discriminate between the essential and spurious terms and how accurately it can determine the coefficients of the essential terms. Since the data were generated using a known model, we know which
terms are essential (those contained in the PDE (4)). If the reference model is unavailable, ensemble regression may be used instead to help distinguish essential terms from spurious ones.

III. RESULTS

As discussed previously, the elements of the library matrix $Q$ are given by the Fourier coefficients of the different terms included in the generalized model (windowed by the envelope $E^{\alpha\beta}(x,t) = (x^2 - 1)^\alpha (t^2 - 1)^\beta$ on each domain $\Omega_k$); hence knowledge of the Fourier spectrum of the data is crucial for an optimal choice of the size of the integration domains $\Omega$ and the weight functions $w_j$. The power spectrum (or, more precisely, the absolute value of the Fourier coefficients) of the noiseless data on the entire physical domain is shown in Figure 1. In space, the spectrum is sharply peaked around a wave number $\kappa \approx 0.625$. At high wave numbers, the spectrum decays exponentially, $P \propto e^{-\kappa/\tilde{\kappa}}$ where $\tilde{\kappa} \approx 0.3$. In time, the spectrum is peaked at zero frequency $\omega$ and decays as a power law, $P \propto \omega^{-\chi}$ with $\chi \approx 2.5$.

Having characterized the data, we turn to the investigation of how the performance of our algorithm depends on the choice of various parameters. Since the number of parameters is quite large, instead of exploring the entire parameter space, we focus on the dependence on one or two parameters at a time, with the remaining parameters staying fixed. Specifically, the noise level $\sigma$ is fixed to 3% and we use the following near-optimal parameters in the sparse regression. The dimensions of the integration domain are $F_x = 2H_x = 14.73$ and $F_t = 2H_t = 75$. This choice corresponds to an equal number of grid points in both directions, $F_x/\delta_x = F_t/\delta_t = 75$. Unless noted otherwise, we use a single set of weights with $\alpha = \beta = 8$, $l = 1$, $m = 2$, and the sparsification parameter is $\gamma = 1.4$. We generally use every combination of 4 weight functions over 50 integration domains, so that the total number of library rows is $K = 200$. To characterize the stochastic effects, for each set of parameters, we used an ensemble of $M = 100$ trials featuring different random distributions of the integration domains and realizations of noise.

First, we tested the ability of the method to reconstruct the correct form of the PDE (4) for various values of $\gamma$ with all other parameters fixed at their near-optimal values. Our iterative regression procedure proved very robust for a fairly wide range of values of $\gamma$. In particular, at a noise level of 30%, it performed perfectly for $1.1 \leq \gamma \leq 2$, with the reconstructed model containing no missing or spurious terms in all of the trials. For the highest noise level considered here (100%), we found perfect performance for $1.2 \leq \gamma \leq 1.5$. In some fraction of the trials, spurious terms appeared at lower $\gamma$ and missing terms at higher $\gamma$, as shown in Figure 2. For reference, without the benefit of the weak formulation, sparse regression failed to correctly reconstruct the lambda-omega reaction-diffusion system, which is only second-order, for noise level as low as 1%.

The accuracy of regression (i.e., model identification) was quantified by computing the relative error in each parameter of the Kuramoto-Sivashinsky equation

$$\Delta c_n = \frac{|c_n - \tilde{c}_n|}{\tilde{c}_n},$$

where $\tilde{c}_n$ and $c_n$ are the true and estimated values of the model parameters, respectively, for $n = 2, 3, 4$. We normalize the estimated parameters so that $c_1 = 1$. In all of the following figures, we plot the estimated mean value of $\Delta c_n$ with 95% error bars, where all of the parameters

![Figure 1](image_url)

FIG. 1. The power spectrum over (a) space and (b) time, normalized so that the maximum is 1. The black dots show the spectrum of the original data. The symbols correspond to the spectra of the windowed data multiplied by envelopes $E^{\alpha\beta}(x,t)$ (with different choices of $\alpha$ or $\beta$) on a “typical” integration domain $\Omega_k$ (i.e., averaged over 1000 uniformly distributed choices of $(x_k,t_k)$). The spatial and temporal frequencies correspond to $\kappa_l = 2\pi l/F_x$ and $\omega_m = 2\pi m/F_t$ for the windowed data.
in the regression procedure are held at their near-optimal values stated previously unless noted otherwise.

In particular, Fig. 3 shows the accuracy of regression as a function of σ for two different choices of data resolution (δs and δt). It is worth noting that the average relative error is ∆cσ ∼ 10^{-10} for all of the parameters for noiseless data with the higher of the two resolutions. However, even for 1% noise, ∆cσ ∼ 2 × 10^{-4}, which is more than three orders of magnitude smaller than what had been achieved in previous studies. The results are very similar for all three parameters; as this is generally the case, in subsequent figures, we only show the generally largest error ∆cα that involves a temporal derivative. We find two distinct regimes. At higher noise levels, the error in evaluating the library matrix entries is due primarily to the averaged effect of noise. Applying the central limit theorem, we find that the relative error scales as

\[ ε_n \sim σ_n \sqrt{\frac{δ_s δ_t}{F_s F_t}}. \]  \hspace{1cm} (15)

At low noise levels, the parameter accuracy is controlled by numerical error, which has two different sources. The first source is a numerical error in the data itself, which is due to the finite accuracy of the integrator that “solves” the Kuramoto-Sivashinsky equation. This source dominates for smaller δs and δt. For experimental data, this source would correspond to systematic error. For larger δs and δt, the parameter inaccuracy is mainly due to the error in computing the library matrix entries based on data that are available on a discrete grid. Suppose we want to use numerical quadratures to evaluate an integral

\[ I = \int_0^L g(x)dx, \]  \hspace{1cm} (16)

where \( g(x) \in C^m \) (i.e., has \( m \) continuous derivatives) and \( g^{(i)}(0) = g^{(i)}(L) \) for all \( 0 \leq i < m \). Then, for the composite trapezoidal rule on a grid with spacing \( h \), the relative error associated with the discretization can be estimated using exact Euler-Maclaurin formulas and is found to scale as \( h^{m+2}|g^{(m+2)}| \) for \( m \) even (or \( h^{m+1}|g^{(m+1)}| \) for \( m \) odd), where a characteristic value of the derivative on the interval \([0, L]\) is used.

Generalizing this result to two dimensions (and assuming \( δ_x ≪ \min(ℓ_x, F_x) \), \( δ_t ≪ \min(ℓ_t, F_t) \)), we find an estimate of the relative discretization error for an element of the library matrix \( Q \) that involves a temporal derivative of order \( ν_t \) and/or spatial derivative of order \( ν_x \):

\[ ε_d \sim \begin{cases} h^{μ+2}, & μ \text{ even} \\ h^{μ+1}, & μ \text{ odd} \end{cases} \]  \hspace{1cm} (17)

where \( h = δ_t/ℓ_t \approx δ_x/ℓ_x \) and \( μ = \min(α − ν_x, β − ν_t) \). It is easy to check that, due to the conditions on \( α \) and \( β \), we always have \( μ ≥ 0 \), as it should be for the trapezoidal rule. The Kuramoto-Sivashinsky equation features terms that all involve derivatives, with the lowest order being one and the highest order four; hence, for even \( α = β ≥ 4 \), the exponent \( μ \) ranges between \( α − 2 \) and \( α + 1 \). Therefore the scaling

\[ ε_d \sim h^{α−2} \]  \hspace{1cm} (18)

dominates for lower \( h \), while the scaling

\[ ε_d \sim h^{α} \]  \hspace{1cm} (19)

dominates for higher \( h \).

The error \( ∆c_α \) can be found using perturbation theory. Let \( ˜Q \) be the library matrix evaluated using a continuous noiseless solution so that \( ˜Q ̄c = 0 \) exactly (we assume that \( ˜Q \) corresponds to the parsimonious model). In the presence of measurement noise and/or discretization error, the error in evaluating each entry \( q_α^{jk} \) of the library

FIG. 2. Fraction \( p \) of identified models with spurious (circles) or missing (squares) terms at maximum noise level (\( σ = 1 \)) as a function of the sparsification parameter \( γ \).

FIG. 3. Parameter errors \( ∆c_α \) as a function of noise level. The circles, triangles, and squares correspond to \( n = 2, 3, 4 \) respectively, and the empty and filled symbols indicate results for data with double resolution (\( δ_x = 0.0982, δ_t = 0.5 \)) and half resolution (\( δ_x = 0.393, δ_t = 2 \)), respectively. The dashed lines show the predicted scaling.
matrix is proportional to $\varepsilon = \max(\varepsilon_d, \varepsilon_n)$, so
\[
Q = \tilde{Q} + \varepsilon \tilde{Q}
\]
for some matrix $\tilde{Q}$ whose entries are distributed as white Gaussian noise. Note that the entries of $Q$ are $O(F_x F_t)$.

The entries of $\tilde{Q}$ have a more complicated scaling that is determined by the Fourier spectrum of the data (i.e., exponential in space, power law in time). Specifically, we find (cf. Fig. 4)
\[
||\tilde{q}_n|| \propto \begin{cases} 
F_x F_t, & \ell_x \ll F_x, F_t \ll \ell_t \\
\varepsilon_n F_x \left( \frac{\ell_x}{F_t} \right)^{\xi_n} \ell_x \ell_t, & F_x \gg \ell_x, F_t \gg \ell_t
\end{cases}
\]
where $\lambda_n = O(\ell_x^{-1})$ and $\xi_n = O(1)$ are some positive constants. To leading order in $\varepsilon$, the least squares solution to (2) is given by
\[
c = \tilde{c} - \varepsilon \tilde{Q}^+ \tilde{Q}\tilde{c},
\]
where $\tilde{Q}^+$ is the Moore-Penrose pseudoinverse of $\tilde{Q}$. Since the elements of $Q$ can be considered uncorrelated, we have for $F_x \gg \ell_x$ and $F_t \gg \ell_t$
\[
\Delta c_n \propto \varepsilon \frac{F_x F_t K^{-1/2}}{\psi(F_x, F_t)},
\]
where the numerator and denominator describe the scaling of the entries of $Q$ and $\tilde{Q}$, respectively. Following from [21],
\[
\psi(F_x, F_t) = e^{-\lambda F_x} \left( \frac{\ell_t}{F_t} \right)^{\xi} \ell_x \ell_t
\]
with some positive constants $\lambda = O(\ell_x^{-1})$ and $\xi = O(1)$.

For low $\sigma$, we have $\varepsilon = \varepsilon_d$ and therefore $\Delta c_n$ is independent of $\sigma$. For high $\sigma$, we have $\varepsilon = \varepsilon_n$, so combining [23] and [15] we find
\[
\Delta c_n \propto \sigma \sqrt{\frac{\delta_d \delta_t}{KF_x F_t \psi(F_x, F_t)}}
\]
The predicted scaling of $\Delta c_n$ with $\sigma$ in both regimes is consistent with the results shown in Fig. [3]. In particular, we find that the effect of changing the resolution of the data is quite minor at high $\sigma$, where $\Delta c_n \propto \varepsilon$ according to (25). At low $\sigma$, the effect is much stronger: for $\alpha = \beta = 8$, we have $\Delta c_n \propto \varepsilon \propto \ell_x^{-4}$ corresponding to (25). The dependence of the scaling in (17) on $\alpha$ and $\beta$ is further confirmed by Fig. 5, which shows results for noiseless data. In the $\alpha = \beta = 4$ case, we observe the scaling law $\Delta c_n \propto \varepsilon \propto \ell_x^{-2}$ in the entire range of $h$ we examined. When $\alpha = \beta = 6$, the parameter error scales according to $\Delta c_n \propto \varepsilon \propto \ell_x^{-4}$ for small $h$ and $\Delta c_n \propto \varepsilon \propto \ell_x^{-1}$ for large $h$, which correspond to the limiting cases [18] and [19], respectively. We should also note that for $h$ as large as $1/4$, the accuracy remains very good. Thus, the method is suitable for fairly sparse data.

As illustrated in Fig. 6, we also observe the scaling for $\Delta c_{n\ell}$ with $K$ predicted by [25]. This scaling is expected to break down when the total area of the integration domains exceeds the area of the physical domain due to the loss of statistical independence between the data on different integration domains, leading to an increased linear dependence of the rows of the library matrix $Q$. We can expect the error to asymptote to
\[
\Delta c_n \propto \varepsilon N_d^{-1/2}
\]
for $K \gg N_d$, where
\[
N_d = \frac{L_x L_t}{F_x F_t}
\]
is the area ratio. For the reference set of parameters, saturation did not occur over the range of $K$ we tested.
To more easily observe the saturation effect, we set \( l = m = 0 \), so that only one weight function is used and the number of integration domains equals \( K \) (rather than \( K/4 \) for nonzero \( l \) and \( m \)). Furthermore, we reduce the size of the physical domain to \( L_x = 16\pi \) and \( L_t = 250 \), so that \( N_d \approx 11 \) is relatively small. As Fig. 6 illustrates, for large \( K \), the parameter accuracy indeed asymptotes to a constant.

The scaling described by (26) can also be observed in the dependence of \( \Delta c_n \) on the size of the physical domain (and hence \( N_d \)) with all other parameters fixed. This dependence is quite important, since it determines how much data needs to be collected to identify the model with meaningful precision. As Fig. 7 illustrates, choosing the physical domain to be just double the size of the (optimal) integration domain in both directions (which corresponds to \( N_d = 4 \)) already yields a rather acceptable accuracy when only one weight function is used. When \( l \) and \( m \) are nonzero, accurate reconstruction is possible even if \( N_d \) is only slightly greater than 1.

Next, we consider how the error in the estimated coefficients depends on the choice of the integration domain size. Figure 8 shows the dependence of the error \( \Delta c_4 \) on the size of the integration domains for two different choices of the weight functions. In panel (a), \( F_t \) is fixed to 75 and \( F_x \) is taken to vary, and in panel (b), \( F_x \) is fixed to 14.73 with \( F_t \) varying. In both cases, we find that there is an optimal domain size with \( F_x \approx 14.73 \) and \( F_t \approx 75 \); moreover, the optimal values remain approximately the same even if we vary the size of the other dimension or the choice of weight functions. For small \( F_x \) and/or \( F_t \), the error is large because (a) the integration domain is too small to effectively average out the influence of noise and (b) the numerical quadrature error is large (both \( c_n \) and \( c_d \) increase as \( F_x \) and/or \( F_t \) decrease). For large \( F_x \) and \( F_t \), we enter the regime described by (23), which predicts that the error should grow exponentially in \( F_x \) and as a power of \( F_t \). Indeed, this is exactly what we observe in Fig. 8. Based on (21), it appears that the optimal choice of \( F_x \) and \( F_t \) corresponds to the crossover between these two regimes, i.e., \( F_x \propto \ell_x \) and \( F_t \propto \ell_t \). Our numerical results suggest that the optimal choice corresponds to \( F_x/\ell_x = F_t/\ell_t \approx 8 \).

Finally, let us address the optimal choice of frequencies appearing in the weight functions (11). Figure 9 shows the effect of varying either \( l \) or \( m \) with all other parameters fixed at their reference values. Specifically, we plot \( \Delta c_4 \) versus \( \omega_l = 2\pi l/F_x \) and \( \omega_m = 2\pi m/F_t \). (Note that when \( l \) or \( m \) is 0, the number of distinct weight functions is halved, so we correspondingly double the number of integration domains to keep the number of rows in the library constant.) One could assume that the optimal
values would be given by the dominant frequencies of the original data (as we discussed previously, the dominant wave number is $\kappa \approx 0.625$ and the dominant temporal frequency is $\omega = 0$). According to Fig. 1, windowing the data broadens the peaks but leaves both dominant frequencies roughly the same: $\kappa \approx 0.8$ ($l = 2$) and $\omega = 0$ ($m = 0$). Unfortunately, it turns out that we cannot use the spectra to exactly predict the optimal frequencies, which are $\kappa \approx 0.4$ ($l = 1$) and $\omega \approx 0.2$ ($m = 2$ or $3$). However, choosing the frequencies based on the spectra still produces reasonably good accuracy (within a factor of 4 or so of the optimal result).

These results suggest that using weight functions with a combination of different frequencies may be more robust and/or accurate. To test this hypothesis, we considered the case in which weight functions with a range of frequencies in space or time were included, with the total number of library rows fixed at 200. However, this approach yielded a decrease in the accuracy, as the broader choice of weight functions did not compensate for a decrease in the number of integration domains. This suggests that the optimal strategy is to use a large number of integration domains while keeping the frequencies of the weight functions fixed.

**IV. CONCLUSIONS**

We have introduced a robust and flexible approach to data-driven discovery of models in the form of nonlinear PDEs. The approach uses a weak formulation, coupled with a novel sparse regression procedure, to obtain a parsimonious description. We have demonstrated its capability to identify PDEs, even with high-order deriv-
tives, from extremely noisy data with unprecedented accuracy. For instance, with 1% noise, we were able to reduce the error in estimating the parameters of the 4th-order Kuramoto-Sivashinsky equation from 50% to just $2 \times 10^{-4}$. Furthermore, whereas correct identification of the functional form of the underlying PDE has been far from guaranteed at any noise level using past approaches, our algorithm was able to reconstruct the Kuramoto-Sivashinsky equation accurately in 100% of cases from data with a signal-to-noise ratio of 100%.

This impressive performance is achieved by shifting the partial derivatives from the data onto a known smooth weight function using integration by parts, thus avoiding the large errors incurred by repeated numerical differentiation. Our method also proved to be well-adapted to sparse data, maintaining errors of less than 0.1% for a grid resolution only 4 times finer than the correlation length/time. Such reliability and high accuracy in the presence of noisy or sparse data is indispensable for analysis of experimental data. Notably, even in the absence of noise, our results compare very favorably with those of previous studies because the discretization error of the algorithm can be made extremely small: for the Kuramoto-Sivashinsky equation, the relative error in all parameters can easily be reduced to $10^{-10}$. It is also important to mention that the computational cost of our algorithm is comparable to that of existing sparse regression methods.

We also derived the scaling laws that describe the accuracy of the regression as a function of the parameters used in the algorithm and the properties of the data. These scaling laws can be used to fully exploit the flexibility of the weak formulation approach by tuning its various parameters. In particular, the size of the input used by the regression can be controlled by choosing both the number of different integration domains and the number of different weight functions. We have shown that the number of integration domains plays a much more important role than the number of weight functions: the best results can be obtained by using a set of weight functions with a fixed shape (frequency and envelope) and a large number of integration domains. Furthermore, we have determined the optimal shape of the weights and the optimal size of the integration domains. The latter turned out to be determined by the correlation length and time describing the data (with the size roughly an order of magnitude larger than these characteristic scales). We have also shown that, although the error can be reduced further by using data on ever-larger physical domains, satisfactory results can be obtained for physical domains that are just a factor of two larger than the optimal integration domain in each dimension.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. CMMI-1725587. DG gratefully acknowledges the support of the Letson Undergraduate Research Scholarship.

V. REFERENCES

1. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Rumbold, C. Theodoropoulos, et al., “Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis,” Communications in Mathematical Sciences 7, 715–762 (2003).
2. J. I. Hsu, X. Gao, S. Sorooshian, and H. V. Gupta, “Precipitation estimation from remotely sensed information using artificial neural networks,” Journal of Applied Meteorology 36, 1176–1190 (1997).
3. M. Raisi, “Deep hidden physics models: Deep learning of nonlinear partial differential equations,” The Journal of Machine Learning Research 19, 932–955 (2018).
4. J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, “Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach,” Physical review letters 120, 024102 (2018).
5. J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz, “On dynamic mode decomposition: theory and applications,” arXiv preprint arXiv:1312.0041 (2013).
6. I. Mezić, “Analysis of fluid flows via spectral properties of the koopman operator,” Annual Review of Fluid Mechanics 45, 357–378 (2013).
7. C. W. Rowley, “Model reduction for fluids, using balanced proper orthogonal decomposition,” International Journal of Bifurcation and Chaos 15, 997–1013 (2005).
8. B. McKeon and A. Sharma, “A critical-layer framework for turbulent pipe flow,” Journal of Fluid Mechanics 658, 336–382 (2010).
9. J. Bongard and H. Lipson, “Automated reverse engineering of nonlinear dynamical systems,” Proceedings of the National Academy of Sciences 104, 9943–9948 (2007).
10. M. Schmidt and H. Lipson, “Distilling free-form natural laws from experimental data,” science 324, 81–85 (2009).
11. D. Xu and O. Khanmohamadi, “Spatiotemporal system reconstruction using fourier spectral operators and structure selection techniques,” Chaos 18, 043122 (2008).
12. O. Khanmohamadi and D. Xu, “Spatiotemporal system identification on nonperiodic domains using chebyshev spectral operators and system reduction algorithms,” Chaos 19, 033117 (2009).
13. L. Brunton, J. L. Proctor, and J. N. Kutz, “Discovering governing equations from data by sparse identification of nonlinear dynamical systems,” Proceedings of the National Academy of Sciences 113, 3932–3937 (2016).
14. 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).
15. P. A. K. Reinbold and R. O. Grigoriev, “Data-driven discovery of partial differential equation models with latent variables,” (2019), https://arxiv.org/abs/1904.04314.
16. 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).
17. M. Bär, R. Hegger, and H. Kantz, “Fitting partial differential equations to space-time dynamics,” Physical Review E 59, 337 (1999).
18. D. P. Vallette, G. Jacobs, and J. P. Colub, “Oscillations and spatiotemporal chaos of one-dimensional fluid fronts,” Physical Review E 55, 4274 (1997).
19. Y. Kuramoto, “Diffusion-induced chaos in reaction systems,” Progress of Theoretical Physics Supplement 64, 346–367 (1978).
20. I. Sivashinsky, “Weak turbulence in periodic flows,” Physica D 17, 243–255 (1985).
L. N. Trefethen and J. Weideman, “The exponentially convergent trapezoidal rule,” SIAM Review 56, 385–458 (2014).