ℓ_1 Regularized pseudo Least Squares based PDE Identification: Recovery Theory

Yuchen He, Xiaoming Huo, Sung Ha Kang, Yajun Mei, Namjoon Suh

February 19, 2020

Abstract

We consider a problem of considerable practical interest: identification of underlying partial differential equation (PDE) from one set of noisy observations. We assume that the governing PDE can be expressed as a linear combination of a subset from a dictionary, containing different linear and nonlinear differential terms. Can we find the correct set of differential terms, which represents the true partial derivate of underlying function with respect to time? In order to answer this question, we use local polynomial approximation in the general framework of ℓ_1-regularized pesudo least squares method. We develop sufficient conditions for the support set recovery in PDE identification problem. Primal-Dual Witness (PDW) construction is employed to achieve the model selection consistency, along with classical conditions in sparse linear regression. In addition, we provide various numerical examples to give intuitions and understanding of these condition in the context of PDE identification.

1 Introduction

Partial Differential equations (PDEs) are widely used to describe many interesting phenomena arising in scientific fields, including physics, social sciences, biomedical sciences, and economics. The forward problem of solving equations or simulating state variables for differential models has been extensively studied. In this paper, we consider an inverse problem in the PDE context, e.g., using the measurements of state variables to estimate the parameters that characterize the system. To be more specific, we assume that the governing PDE is a linear combination of a subset of a prescribed dictionary containing different differential terms, and the objective is to identify PDE or to find the correct set of coefficients.

From the methodology or algorithm point of view, we develop a two stage method for PDE Identification based on the ℓ_1-regularized Pseudo Least Square (ℓ_1-PsL) model and local polynomial approximation. In the first stage, from given noisy observations, we propose to estimate the underlying function and its derivatives via local polynomial fitting. In the second stage, we propose to estimate model parameters via ℓ_1 regularized least-squares methods. We note that this two-stage method has been applied in the ordinary differential equations (ODE) setting. Liang and Wu [1] established the consistency and asymptotic normality of the pseudo-least square estimator in ODE setting, where they used Local Polynomial Regression to estimate the state variables under the noisy data. Similarly, Chen and Wu [2, 3] studied the parameter estimation of ODE models with varying coefficients. However, research is limited when using this two-stage method to the PDE context.

Reasons for using the local polynomial fitting to the estimation of state variables and their derivatives in PDE are mainly three-folded. First, the core idea of many numerical finite difference approximation is based on the polynomial approximation of the underlying continuous function. For noisy data it is reasonable to consider a local polynomial fitting for approximation with denoising effects. Secondly, local polynomial fitting enables function fitting and direct derivative estimating at the same time, thanks to construction of the optimization problem via Taylor expansion [4]. Additionally, rich literatures in asymptotic properties and uniform convergence of the estimator [4, 5, 6] allow us to explore the tail-bound behavior of the truncation error.

It has been widely used to apply ℓ_1 regularized pseudo least square method to perform a variable selection. Hayden [7] studied the model selection problem via LASSO under PDE context. The author empirically showed that the method works well in various important equations such as Burger's equation, Navier-Stokes equation, Swift-Hohenberg equation, just to name a few. Kang, Liao, and Liu [8] considered PDE identification problem using numerical time evolution. The authors utilized LASSO to select candidate monomials, then proposed the time evolution error to select the underlying true model. Although these works demonstrated some empirical success of the LASSO in PDE, a rigorous theoretical justification on the usage of LASSO still remains vague in PDE identification.

From the theoretical point of view, our paper is the first work to provide the theoretical justification of ℓ_1 LASSO with local polynomial approximation in PDE identification problem. Our main theoretical contribution is to establish a reasonable sufficient condition for the support recovery of LASSO in PDE identification setting, thereby bridging the gap between practice and theory. Our main tool is Wainwright's Primal-Dual Witness (PDW) construction [9], which is a popular technique to prove model selection consistency in statistics and machine learning , see [10, 11, 12, 13, 14, 15]. One essential assumption of LASSO is a Mutual Incoherence Condition, which states that the large number of irrelevant predictors cannot exhibit an overly strong influence on the subset of relevant predictors. This condition arose from high-dimensional setting where "p" is much larger than "n." Here we express this condition explicitly in the PDE context, and demonstrate that it is reasonable for PDE identification.

It is not trivial to establish the theoretical properties of our proposed method of using LASSO with local polynomial approximation in PDE identification. There are a couple of critical challenges. First, the errors (i.e., measurement errors) under our setting are not i.i.d., but dependent. In the statistics literature, when the noise is assumed to be i.i.d. sub-gaussian, the tail bound of the term related with the noise can be easily shown to have an exponential decay via a combination of Chernoff and Union bound, and thus one is able to establish the strict dual feasibility in PDW construction [9]. However, in the setting of this paper, the error term, which involves several different sources of errors, is neither mean 0 nor i.i.d. Second, all columns of the feature matrix are estimated via local polynomial regression, and thus are random. As a result, we need to construct Mutual Incoherence Assumption that will be held on the estimated feature matrix with high probability.

The rest of the paper is organized as follows. In Section 2, we present the model formulation for PDE identification. We propose a ℓ_1-regularized pseudo-least square method and introduce a local polynomial regression for estimating derivatives. In Section 3, we present our main theoretical results, and provide three assumptions to achieve a model selection consistency in PDE identification problem. We connect these assumptions to PDE identification setting in Section 4, and present various numerical results of the proposed model in Section 5. We conclude the article with a discussion in Section 6. The detailed technical proofs of the presented theorems are given in the Appendix.

2 Proposed Method for PDE Identification

Let $u : \mathbb{R} \times [0, +\infty) \to \mathbb{R}$ be a real-valued function, x to be the spatial and t to be the temporal variables. Suppose that within a bounded region $\mathbb{R} \times [0, \infty)$, u satisfies an evolutionary partial differential equation (PDE):

$$u_t(x, t) = F(u, u_x, u_{xx}, \dots)\ ,\ (x, t) \text{ in } \Omega \subset \mathbb{R} \times [0, +\infty).$$

Here, u_t (or $\partial_t u$) denotes the partial derivative of u with respect to t; for $p = 0, 1, 2, \ldots$, $\partial_x^p u$ denotes the p-th order partial derivative of u with respect to x; F is a real-valued mapping and Ω is a bounded open subset of the time-space domain. We take $\Omega = (0, X_{\max}) \times (0, T_{\max})$ for some finite positive numbers $0 < X_{\max}, T_{\max} < +\infty$ and assume that F is a degree 2 polynomial:

$$u_t(x,t) = F(u, u_x, u_{xx} \ldots; \boldsymbol{\beta}^*) := \beta_0^* + \beta_1^* u + \beta_2^* u_x + \beta_3^* u_{xx} + \cdots + \beta_{p,q}^* \partial_x^p u \partial_x^q u + \ldots\ ,$$
$$(x,t) \text{ in } \Omega = (0, X_{\max}) \times (0, T_{\max})\ . \tag{1}$$

Let the coefficient vector be $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \ldots, \beta_{p,q}^*, \ldots)$ with real entries. We call the monomials in the right hand side in (1) as *feature variables.* This format encloses various important classes of PDEs, e.g., advection-diffusion-decay equation characterizing pollutant distribution in fluid, Burgers' equation modeling the traffic flow, Kolmogorov-Petrovsky-Piskunov equation describing phase transitions [16], and Korteweg-de Vries equation simulating the shallow water dynamics [17], and others. We set a finite integer upper-bound, $P_{\max} > 0$, for the possible orders of the partial derivatives of u with respect to x in (1). Hence we assume that $\boldsymbol{\beta}^* \in \mathbb{R}^K$, with $K = 1 + 2(P_{\max} + 1) + \binom{P_{\max}+1}{2}$, so that constant and any term of the form $\partial_x^p u$ or $\partial_x^p u \partial_x^q u$, $0 \le p, q \le P_{\max}$, are contained in (1). Many coefficients may be zero. We denote $\mathcal{S}(\boldsymbol{\beta}^*)$, or simply $\mathcal{S}$, as the support of the coefficient vector $\boldsymbol{\beta}^*$, the set of indices of the non-zero entries.

The given data set $\mathcal{D} = \{(X_i, t_n, U_i^n)|\ i = 0, 1, \ldots, M-1;\ n = 0, 1, \ldots, N-1\} \subseteq \Omega \times \mathbb{R}$ consists of $M \times N$ data, $M, N \in \mathbb{N}$, $N, M \ge 1$. Here $(X_i, t_n) \in \Omega$ represents (structured or unstructured) space-time sample points, and U_i^n is a representation of $u(X_i, t_n)$ contaminated by additive Gaussian noise:

$$U_i^n = u(X_i, t_n) + \nu_i^n,\ \ \nu_i^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)\ .$$

Here $\mathcal{N}(0, \sigma^2)$ denotes the centered normal distribution with variance $\sigma^2 > 0$. We propose to identify the coefficient vector $\boldsymbol{\beta}^*$ from this given set $\mathcal{D}$. As a basic requirement, the mean of the noisy data is uniformly bounded, i.e., for any integer $N \ge 1$ and $M \ge 1$, $\max_{i=0,\cdots,M-1; n=0,\cdots,N-1} E|U_i^n|^s \le C_s < \infty$ for some constant $C_s > 0$ and an integer $s \in \mathbb{N}$.

Throughout this paper, we write bold lower-case letters for vectors and bold upper-case letters for matrices. We use $||\cdot||_1$, $||\cdot||_2$, and $||\cdot||_\infty$ to denote the 1-norm, 2-norm, and ∞-norm respectively, of a matrix or a vector. For a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$, $||\mathbf{A}||_1 = \max\limits_{j=1,\ldots,m} \sum_{i=1,\ldots,n} |A_{i,j}|$, $||\mathbf{A}||_2 =$ maximal singular value, and $||\mathbf{A}||_\infty = \max\limits_{i=1,\ldots,n} \sum_{j=1,\ldots,m} |A_{i,j}|$; for a vector $\mathbf{v} \in \mathbb{R}^n$, $||\mathbf{v}||_1 = \sum\limits_{i=1,\ldots,n} |v_i|$, $||\mathbf{v}||_2 = \sqrt{\sum\limits_{i=1,\ldots,n} v_i^2}$, and $||\mathbf{v}||_\infty = \max\limits_{i=1,\ldots,n} |v_i|$. We use $|\mathcal{I}|$ to denote the number of elements in a set $\mathcal{I}$. For an index set $\mathcal{I} \subseteq \{1, \ldots, m\}$, we use $\mathbf{A}_{\mathcal{I}} \in \mathbb{R}^{n \times |\mathcal{I}|}$ to denote a matrix obtained by taking the columns of $\mathbf{A}$ indexed by $\mathcal{I}$.

2.1 ℓ_1-regularized Pseudo Least Square Model

We propose to estimate the coefficient vector $\boldsymbol{\beta}^* \in \mathbb{R}^K$ of the PDE (1), using a minimizer $\widehat{\boldsymbol{\beta}}^\lambda$ of an ℓ_1-regularized pseudo least square (ℓ_1-PsL) problem:

$$\widehat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \frac{1}{2NM} \sum_{i=0}^{N-1} \sum_{n=0}^{M-1} \left(\widehat{\partial_t u_i^n} - F(\widehat{u_i^n}, \widehat{(\partial_x u)_i^n}, \ldots, \widehat{(\partial_x^{P_{\max}} u)_i^n}; \boldsymbol{\beta})\right)^2 + \lambda ||\boldsymbol{\beta}||_1\ . \tag{2}$$

Here, $\widehat{(u_t)_i^n}$ and $\widehat{(\partial_x^p u)_i^n}$, $p = 0, 1, \ldots, P_{\max}$, are smooth estimators for $(u_t)_i^n$ and $(\partial_x^p u)_i^n = \partial_x^p u(X_i, t_n)$ respectively derived from the data $\mathcal{D}$. The details of the estimators are presented in subsection 2.2. The parameter $\lambda > 0$ is a penalty parameter that may depend on the data size N and M. The first term of (2) imposes the requirement that the estimated time derivatives and space derivatives are related via a polynomial approximation. The second term is an ℓ_1-norm regularizer which encourages the sparsity in the recovered coefficient vector $\widehat{\boldsymbol{\beta}}^\lambda$. The word *pseudo* comes from the fact that the conventional assumption of independence among the residues is violated. If $\lambda = 0$, the proposed ℓ_1-PsL model (2) reduces to the PsL model introduced in [1] for estimating the ODE coefficients with known support.

We introduce matrix notations for compact expressions. We let $\mathbf{u}_t \in \mathbb{R}^{NM}$ denote the vectorization of $(u_t(X_i, t_n))_{i,n}$ in a dictionary order prioritizing the spatial dimension; that is, $\mathbf{u}_t^T = \begin{bmatrix} u_t(X_0, t_0) & u_t(X_1, t_0) & \cdots \end{bmatrix}$. Define the *feature matrix*, $\mathbf{F} \in \mathbb{R}^{NM \times K}$, as the collection of values of feature variables organized as follows:

$$\mathbf{F} = \begin{bmatrix} 1 & u(X_0,t_0) & \partial_x u(X_0,t_0) & \cdots & \partial_x^p u(X_0,t_0)\partial_x^q u(X_0,t_0) & \cdots \\ 1 & u(X_1,t_0) & \partial_x u(X_1,t_0) & \cdots & \partial_x^p u(X_1,t_0)\partial_x^q u(X_1,t_0) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 1 & u(X_{M-1},t_0) & \partial_x u(X_{M-1},t_0) & \cdots & \partial_x^p u(X_{M-1},t_0)\partial_x^q u(X_{M-1},t_0) & \cdots \\ 1 & u(X_0,t_1) & \partial_x u(X_0,t_1) & \cdots & \partial_x^p u(X_0,t_1)\partial_x^q u(X_0,t_1) & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots \\ 1 & u(X_{M-1},t_{N-1}) & \partial_x u(X_{M-1},t_{N-1}) & \cdots & \partial_x^p u(X_{M-1},t_{N-1})\partial_x^q u(X_{M-1},t_{N-1}) & \cdots \end{bmatrix} .$$

We obtain $\widehat{\mathbf{u}}_t \in \mathbb{R}^{NM}$ and $\widehat{\mathbf{F}} \in \mathbb{R}^{NM \times K}$ by replacing the entries of $\mathbf{u}_t$ and $\mathbf{F}$ respectively with the corresponding estimators, i.e., $\widehat{(u_t)_i^n}$ and $\widehat{(\partial_x^p u)_i^n}$ (subsection 2.2). Applying these notations, ℓ_1-PsL (2) can be rewritten in the following matrix form:

$$\widehat{\boldsymbol{\beta}}^\lambda = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^K} \frac{1}{2NM} (\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta})^T (\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta}) + \lambda ||\boldsymbol{\beta}||_1 \ . \tag{3}$$

Observe that (3) is formally identical to LASSO for high-dimensional sparsity recovery: $\widehat{\mathbf{u}}_t$ is the response vector, and $\widehat{\mathbf{F}}$ corresponds to the design matrix.

The fact that the data is generated from a solution of a PDE provides a unique structure for the feature matrix and a special connection between residue minimization and numerical consistency. We define the *PDE estimation error*, $\boldsymbol{\tau} \in \mathbb{R}^{NM}$, as the remainder of the underlying PDE (1) with the time derivatives and the feature variables substituted by corresponding estimators:

$$\boldsymbol{\tau} = \widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta}^* \ . \tag{4}$$

By Cauchy-Schwarz and Minkowski inequality, the first term of (3) is bounded from above by

$$\frac{1}{2NM}\big(||\boldsymbol{\tau}||_2 + ||\widehat{\mathbf{F}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)||_2\big)^2 \leq \frac{1}{2}\big(||\boldsymbol{\tau}||_\infty + ||\widehat{\mathbf{F}}_{\mathcal{S}}||_\infty ||\boldsymbol{\beta}_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}||_\infty + ||\widehat{\mathbf{F}}_{\mathcal{S}^c}||_\infty ||\boldsymbol{\beta}_{\mathcal{S}^c}||_\infty\big)^2 \ . \tag{5}$$

This upper-bound decomposition suggests that, in order to control residual errors resulted from the data regression, it is sufficient to minimize the PDE estimation error $||\boldsymbol{\tau}||_\infty$ coming from the PDE discretization, and the coefficient error $||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_\infty$ coming from misspecification of the feature variables. As suggested in (5), the coefficient error consists of two parts: comparison between the entries of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ at the positions indexed by $\mathcal{S}$, as well as the vanishing of $\boldsymbol{\beta}$ at the those indexed by $\mathcal{S}^c$. Since $\mathcal{S}$ is unknown a priori, introducing the ℓ_1-regularizer helps to filter the correct feature variables while enforcing the other coefficients null.

2.2 Local Polynomial Regression Estimators for Derivatives

We apply the local polynomial regression [4] to obtain $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$. This approach is advantageous in many aspects. Asymptotically, there are theoretical guarantees about the bias and variance of the estimators, and it is proved that the estimation has asymptotic minimax efficiency [4]. One can simultaneously estimate u and its various partial derivatives at any point via one single regression. Moreover, the core idea of common finite difference approaches is based on the local polynomial approximation, thus for noisy data we extend to a local polynomial fitting which can give denoising effect and stabilize the process.

Given the data $\{(X_i, t_n, U_i^n)\}$ with $i = 0, 1, \ldots, M-1$ and time domain $n = 0, 1, \ldots, N-1$, we consider two kinds of local polynomial regressions. One is to locally fit a degree 4 polynomial over to obtain $\widehat{u}_t(X_i, \cdot)$ for each fixed space point X_i. The other is to locally fit a degree $p+3$ polynomial to $\widehat{u}_t(x, t_n)$ so as to compute its p-th order derivative $\widehat{\partial_x^p u}(\cdot, t_n)$ at each fixed time point t_n, for each $p = 0, 1, \ldots, P_{\max}$. This allows us to transform the task of approximating a function defined over a two-dimensional domain to a series of one-dimensional local polynomial regressions.

To be more specific, we solve the following $M + (1 + P_{\max})N$ optimization problems for local polynomial fittings:

$$
\begin{aligned}
&(\widehat{b}_j(X_i, t))_{j=0,1,\ldots,4} = \arg \min_{b_j(t)\in\mathbb{R}, j=0,1,\ldots,4} \sum_{n=0}^{N-1} \Big(U_i^n - \sum_{j=0}^{4} b_j(t)(t_n - t)^j\Big)^2 \mathcal{K}\Big(\frac{t_n - t}{h_N}\Big)\ , \\
&\qquad \text{for } i = 0, 1, \ldots, M-1\ ; \\
&(\widehat{c}_j^p(x, t_n))_{j=0,1,\ldots,p+3} = \arg \min_{c_j(t)\in\mathbb{R}, j=0,1,\ldots,p+3} \sum_{n=0}^{N-1} \Big(U_i^n - \sum_{j=0}^{p+3} c_j^p(t)(X_i - x)^j\Big)^2 \mathcal{K}\Big(\frac{X_i - x}{w_{p,N}}\Big)\ , \\
&\qquad \text{for } n = 0, 1, \ldots, N-1 \text{ and } p = 0, 1, \ldots, P_{\max}.
\end{aligned}
$$

Next, we then set $\widehat{u}_t(X_i, t) = \widehat{b}_1(X_i, t)$ and $\widehat{\partial_x^p u}(X_i, t) = p!\widehat{c}_p^p(x, t_n)$. They are used to assemble $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$. We choose $\mathcal{K}$ to be the Epanechnikov kernel defined by:

$$
\mathcal{K}(z) = \frac{3}{4}(1 - z^2)_+\ ,\ z \in \mathbb{R}\ , \tag{6}
$$

where $(\cdot)_+$ refers to taking the positive part; and h_N and $w_{p,N}$ are the bandwidth parameters. It is shown in [4] that, the Epanechnikov kernel (6) has optimal performance at interior points and nearly optimal at the most boundary points. It is easy to verify that $\mathcal{K}$ is uniformly continuous; absolutely integrable with respect to Lebesgue measure on the line; $\mathcal{K}(z) \to 0$ as $|z| \to +\infty$; and $\int |z \log |z||^{1/2} |dK(z)| < +\infty$.

In Figure 1, we compare the estimated partial derivatives of $u(x, t) = 2\sin(\pi(x - 2.5t)/4)$, for $0 \le x \le 1$ and $0 \le t \le 2$, using noisy data ($\sigma = 0.01$) with those computed directly using forward difference. Although the noise on the function u is visually harmless in (a), we see in (b) and (c) that the error is magnified through the forward difference scheme. Applying the local polynomial regression clearly amends the estimation for the points.

3 Recovery Theory for ℓ_1-PsL based PDE Identification

We present our main results on the ℓ_1-PsL model for PDE identification. We focus on four aspects of the minimizer $\widehat{\boldsymbol{\beta}}^\lambda$ of (2): (i) uniqueness, (ii) proper support recovery ($\widehat{\boldsymbol{\beta}}^\lambda \subseteq \boldsymbol{\beta}^*$), (iii) ℓ_∞-norm of the coefficient error on true support ($||\widehat{\boldsymbol{\beta}}_m S^\lambda - \boldsymbol{\beta}_m S^*||_\infty$), and (iv) signed support recovery ($\text{sgn}\, \widehat{\boldsymbol{\beta}}^\lambda = \text{sgn}\, \boldsymbol{\beta}^*$). These properties strongly affect the behaviors of the identified PDE compared to those of the underlying PDE. Under some assumptions, we show that the PDE identified by the ℓ_1-PsL model converges to the true PDE with high probability.

Here is an outline of proving our main theorem: Using the primal-dual witness method [9], we construct the pair $(\check{\boldsymbol{\beta}}, \check{\mathbf{z}}) \in \mathbb{R}^K \times \mathbb{R}^K$ with $\check{\boldsymbol{\beta}}$ satisfying the KKT equation associated with (2) such that $\mathcal{S}(\check{\boldsymbol{\beta}}) = \mathcal{S}(\boldsymbol{\beta}^*)$, and $\check{\mathbf{z}}$ being a subgradient of the ℓ_1-norm evaluated at $\check{\boldsymbol{\beta}}_{\mathcal{S}}$. The desired properties of $\widehat{\boldsymbol{\beta}}^\lambda$ hold immediately if $||\check{\mathbf{z}}||_\infty < 1$ with high probability, based on Lemma 2 and 3 of [9]. We show that bounding the norm of the PDE estimation error $||\boldsymbol{\tau}||_\infty$ is sufficient. This

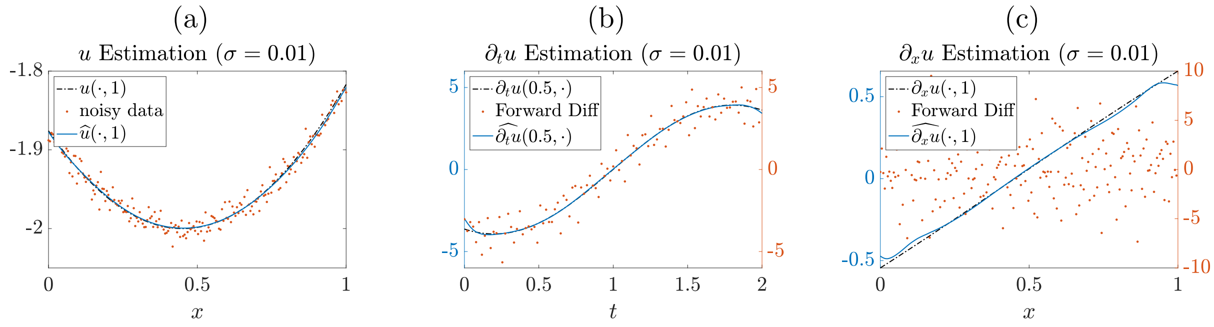

Figure 1: Local polynomial regression estimation for $u(x,t) = 2\sin(\pi(x-2.5t)/4)$, for $0 \le x \le 1$ and $0 \le t \le 2$, and its derivatives using noisy data ($\sigma = 0.01$). (a) u versus $\widehat{u}$ when $t = 1$. (b) u_t versus $\widehat{u}_t$ when $x = 0.5$ (c) $\partial_x u_x$ versus $\widehat{\partial_x u}$ when $t = 1$. $\Delta x = 1/(M-1)$ and $\Delta t = 1/(N-1)$, where $M = 200$ and $N = M^{7/8} \approx 103$; $h_N = N^{-1/7} \approx 0.5158$ and $w_M = 1.2M^{-1/8} \approx 0.6188$.

quantity is further associated with two important aspects of the local polynomial regression estimation: asymptotic bias [4] and asymptotic uniform convergence [6]. With well-designed bandwidths for the kernel of the local polynomial regression, we complete the proof. We first list the important assumptions in Section 3.1, followed by the main theorem in Section 3.2.

3.1 Model Assumptions and Implications on PDEs

We introduce three key conditions whose deterministic versions are frequently assumed to hold in ℓ_1-regularized regression models. They were typically used to prove sufficient conditions for exact sparse recovery. In our context, they carry special meanings and impose general properties on the observed solution u and the underlying PDEs for a successful PDE identification from the given data. In the next section, we will provide more detailed discussions on these three key conditions, and provide deeper insights under in the PDE setting.

Invertibility condition — Possibility of identification of PDE from a single solution

$$\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}} \text{ is invertible, almost surely.} \tag{A1}$$

This assumption is traditionally related to the uniqueness of the solution of a linear regression model. If it fails, multicolinearity exists among the columns of $\widehat{\mathbf{F}}_{\mathcal{S}}$, which leads to intrinsic ambiguity in the modeling. The necessity of (A1) largely depends the fact that, in many practical settings, e.g., ocean surface monitoring [18] and flock tracking [19], one may not be able to record multiple solutions of a single PDE, determined by some unique combination of numerous known and unknown environmental factors. For example, it is impossible to choose between $u_t = 3u_{xx} + u_x$ and $u_t = 5u_{xx} + u_x$ when we only observe $u(x,t) = x + t$. In Section 4.1, we prove that, provided with sufficiently many data, (A1) is equivalent to whether u is a common solution to the underlying PDE and a steady-state equation of the PDE.

Mutual incoherence condition — Exhibition of signature variation For some *incoherence parameter* $\mu \in (0,1]$ and $P_\mu \in [0,1]$:

$$\mathbb{P}\big[||\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}})^{-1}||_\infty \le 1-\mu\big] \ge P_\mu\,. \tag{A2}$$

Compared to the invertibility condition (A1), (A2) is quantitative and more challenging to verify. This condition sets apart the group of correct feature variables from the the group of the others with probability P_μ. A larger μ indicates less similarity between these two groups, thus it is easier to identify the correct model. In the literature, the deterministic version of (A2) is called the *mutual incoherence condition* [20, 21, 9], and we refer the readers to [22] for more well-known recoverability conditions and discussion on their relations.

The validity of (A2) is closely related to the characteristic shapes or movements of the solution u. The observed solution u of the underlying PDE should present the signature variation within the time-space domain for us to distinguish the correct feature variables from the others.

Minimal eigenvalue condition —Threshold on noticeable magnitude There exists some constant $C_{\min} > 0$ such that:

$$\Lambda_{\min}\Big(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\Big) \geq C_{\min} \text{ , almost surely.} \tag{A3}$$

Here $\Lambda_{\min}(\mathbf{A})$ denotes the minimal eigenvalue of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$. This condition can be considered as strengthend (A1). Similarly to (A2), (A3) concerns the quantitative properties of the feature variables associated with the solution u. However, (A3) generally does not involve the characteristic variation of u that is unique to the combination of the underlying feature variables.

3.2 The Main Result: Uniqueness and Proper Support Recovery

Theorem 3.1. *Provided with $\mathcal{D} = \{(X_i, t_n, U_i^n) \mid i = 0, 1, \cdots, M-1,\ n = 0, 1, \cdots, N-1\} \subset \Omega$ and under the assumptions in Section 3.1, there exists a constant $C' > 0$ independent of N and M, such that if we take $M = N^{(6+P_{\max})/7}$, $h_N = N^{-1/7}$ in the time direction, $w_M = M^{-1/(6+P_{\max})}$ in the space direction, and*

$$\lambda \geq \frac{C'}{\mu}\sqrt{\frac{K(6+P_{\max})\ln N}{7N^{4/7}}}\ , \tag{7}$$

then with probability greater than

$$P_\mu - \underbrace{(8K+2)N^{(13+P_{\max})/7}K\exp\big(-\frac{(N^{1/7}-||u||_{L^\infty(\Omega)})^2}{2\sigma^2}\big)}_{:=P'(K,N,\sigma,||u||_{L^\infty(\Omega)})\ \text{or simply}\ P'} \longrightarrow P_\mu\ ,\ \text{as } N\to\infty\ , \tag{8}$$

the minimizer $\widehat{\boldsymbol{\beta}}^\lambda$ of (2) is unique, and its support is properly contained in the true support:

$$\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)\ .$$

Proof. See Appendix A □

In Theorem 3.1, we apply the PDW [9] to arrive at the necessary conditions for the PDW dual variable $\widetilde{\mathbf{z}}$ to be a local minimizer. Whether the support of $\widehat{\beta}^\lambda$ is contained in the true support depends on the magnitude of $||\check{\mathbf{z}}_{\mathcal{S}^c}||_\infty$. This involves a careful estimation of the error $\boldsymbol{\tau}$ defined in (4), which is a random vector with inaccessible probability distribution. To overcome this challenge, we imply the results in [4] and [6], and derive an explicit upper-bound for $||\boldsymbol{\tau}||_\infty$ as $N, M \to \infty$ in probability sense. This strategy is similarly applied in [1] to prove the asymptotic convergence of the PsL estimators for dynamical systems.

The threshold for λ characterized by (7) shows that when the number of data increases, there is more flexibility in tuning this parameter. If the incoherence parameter μ is small, or equivalently, the group of correct feature variables and the group of the others are similar, then the threshold (7) increases. In Theorem 1 of [9], the threshold for the regularization parameter shows consistent behavior.

Note that, $P_\mu - P'$ in (8) is truly a probability for sufficiently large N, and P' reduces to 0 exponentially fast after certain amount of data is collected. Figure 2 shows the dependence of

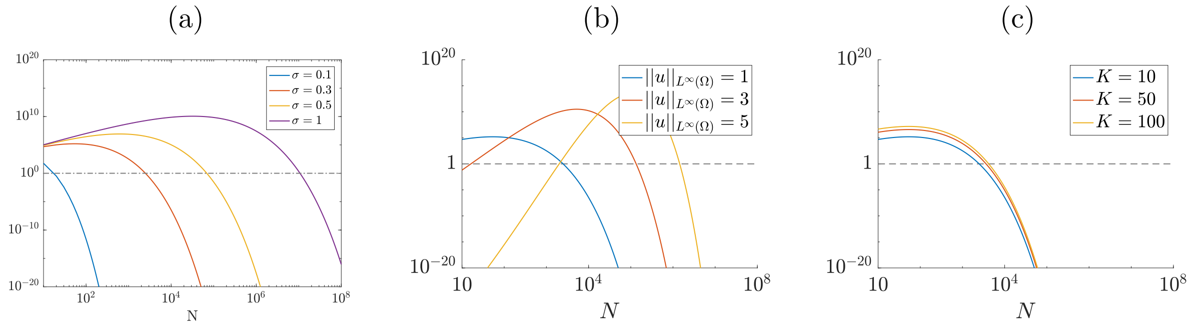

Figure 2: As N increases, the variation of $P'(K, N, \sigma, ||u||_{L^\infty(\Omega)})$ for different (a) noise level σ ($||u||_{L^\infty(\Omega)} = 1$, $K = 10$); (b) solution bound $||u||_{L^\infty(\Omega)}$ ($\sigma = 0.3$, $K = 10$); and (c) number of candidate feature variables K ($\sigma = 0.3$, $||u||_{L^\infty(\Omega)} = 3$). In all cases, $P_{\max} = 2$.

P' on the noise level σ, solution bound $||u||_{L^\infty(\Omega)}$, and the number of candidate feature variables K. As seen in (a), when the underlying data is contaminated by heavier noise, it requires more data to guarantee the conclusion of Theorem 3.1 with high probability. (b) shows that with higher function magnitude, the range for P' exceeding 1 becomes narrower. There is little effect of K on the effectiveness of (8) as shown in (c).

Unlike Theorem 1 of [9], the uniqueness and proper support recovery in our case hold only up to probability $P_\mu \in [0, 1]$, rather than 1 as $N \to \infty$. The limiting probability P_μ is determined by the mutual incoherence condition (A2); therefore, for different combinations of the observe solution and the underlying PDE, the certainty about the conclusion in Theorem 3.1 can vary. We numerically study P_μ for various examples in Section 5 which demonstrate this complexity.

We also observe that the highest partial derivative order of the candidate feature variables, $P_{\max}$, has effect on the convergence rate of (8). The accuracy of high-order derivative estimation is deprecated using local polynomial regression. We often need to set $P_{\max}$ large enough to lower the model bias. The larger $P_{\max}$ indicates more data is needed before P' starts to exponentially drop to 0.

3.3 ℓ_∞ Error Bound and Signed Support Recovery

From (22), let $\Delta \mathbf{u}_t = \widehat{\mathbf{u}}_t - \mathbf{u}_t$, $\Delta \mathbf{F} = \widehat{\mathbf{F}} - \mathbf{F}$ denote the error terms, then we have:

$$\boldsymbol{\beta} - \boldsymbol{\beta}^* = \left(\widehat{\mathbf{F}}^T\widehat{\mathbf{F}}\right)^{-1}\left[\widehat{\mathbf{F}}^T(\Delta\mathbf{u}_t - \Delta\mathbf{F}\boldsymbol{\beta}^*) - \lambda N M \mathbf{z}\right].$$

Focusing on the indices corresponding to the support of $\boldsymbol{\beta}^*$, the error of the coefficient estimation is bounded

$$\begin{aligned}
\max_{k\in\mathcal{S}} |\boldsymbol{\beta}_k - \boldsymbol{\beta}^*_k| &\le ||\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}||_\infty ||\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}||_\infty + \lambda N M ||\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}||_\infty \\
&\le ||\left(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}/(NM)\right)^{-1}||_\infty \left(||\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}||_\infty/(NM) + \lambda\right) \\
&\le \sqrt{K} C_{\min} \left(||\widehat{\mathbf{F}}_{\mathcal{S}}^T\boldsymbol{\tau}||_\infty/(NM) + \lambda\right) \\
&\le \sqrt{K} C_{\min} \left((||\mathbf{F}_{\mathcal{S}}||_1 + ||\Delta\mathbf{F}_{\mathcal{S}}||_1)\frac{||\boldsymbol{\tau}||_\infty}{NM} + \lambda\right) \\
&\le \sqrt{K} C_{\min} \left((||\mathbf{F}_{\mathcal{S}}||_1 + NM\sqrt{K}||\Delta\mathbf{F}||_{\max})\frac{||\boldsymbol{\tau}||_\infty}{NM} + \lambda\right) \\
&= \sqrt{K} C_{\min} \left(\frac{||\mathbf{F}_{\mathcal{S}}||_1 ||\boldsymbol{\tau}||_\infty}{NM} + \sqrt{K}||\Delta\mathbf{F}||_{\max}||\boldsymbol{\tau}||_\infty + \lambda\right).
\end{aligned}$$

Hence, applying Theorem A.1 and Lemma 3 (b) of [9] gives the following result.

Theorem 3.2. *Suppose the conditions for Theorem 3.1 hold, with probability greater than*

$$P_\mu - (8K+2)N^{(13+P_{\max})/7} K \exp\Big(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}\Big) \longrightarrow P_\mu \ , \ \text{as } N \to \infty \ ,$$

then

$$||\widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}||_\infty \leq \sqrt{K} C_{\min} \left(\frac{C'\big(||\mathbf{F}_{\mathcal{S}}||_1 + C'\sqrt{K}\big)}{N^{(19+P_{\max})/7}} + \lambda \right) . \tag{9}$$

Moreover, if $\min_{k\in\mathcal{S}} |\boldsymbol{\beta}^*_k| > \sqrt{K} C_{\min} \left(\frac{C'(||\mathbf{F}_{\mathcal{S}}||_1 + C'\sqrt{K})}{N^{(19+P_{\max})/7}} + \lambda \right)$*, then* $\widehat{\boldsymbol{\beta}}^\lambda$ *has the correct signed support.*

The upper-bound for the ℓ_∞-norm of the coefficient error in (9) consists of two components. The first one contains information about the underlying PDE, the data size, and the number of candidate features. As N increases to ∞, this part converges to 0 without explicit dependence on the choice of feature variables selected from ℓ_1-PsL. The second component is simple: $\sqrt{K} C_{\min}\lambda$. When N increases, this part does not vary. We note that by adjusting λ so that it is above the threshold stated in Theorem 3.1, ℓ_1-PsL is guaranteed to recover the correct feature variables under some conditions. Since the threshold (7) decreases to 0 as $N \to \infty$, we see that we can choose smaller λ to achieve the recovery. As an overall effect, increasing the data size N leads to more accurate coefficient estimation.

An important implication of Theorem 3.2 is the correct signed support. Many PDEs are sensitive to the sign of the coefficients. For example, changing the sign of the advection term in transport equation reverses the moving direction, and changing the sign of the Laplacian term of the heat equation leads to instability. Theorem 3.2 guarantees the correct signs provided that the magnitudes of the coefficients of the correct feature variables are larger than a threshold same as (9). Asymptotically, the recovered coefficients are close to the true ones, and the signs are correct even for those with small absolute values.

4 Recoverability Theory and PDE Identification

We connect the three key conditions used in Theorem 3.1 to the setting of identification of PDE. First, we examine the invertibility condition (A1) in connection to the accuracy of the system. In particular, even with sufficiently many data, we note that when the underlying PDE and a steady-state equation have a common solution, the condition (A1) is invalid. Yet, this is relatively uncommon as illustrated by examples in this section. A couple of examples are presented to give insight into the conditions (A2) and (A3).

4.1 Invertibility Condition (A1)

When there exist sufficiently many data, we prove that (A1) fails if and only if the observed function u is a common solution of the underlying PDE and a steady-state equation. The steady-state equation is a stationary PDE, $u_t = 0$, of the same type as the underlying one, i.e., the coefficient vectors for the feature variables share the same support. First, we prove the following technical lemma:

Lemma 4.1. *Denoting* $\mathbf{F}_p$ *as the column of the feature matrix corresponding to the feature* $\partial_x^p u$*,* $p = 0, 1, \cdots, P_{\max}$*, then*

$$\left| \int_\Omega \partial_x^p u(x,t)\, dx\, dt - \mathbf{1}^T \mathbf{F}_p \Delta x \Delta t \right| \to 0 \ ,$$

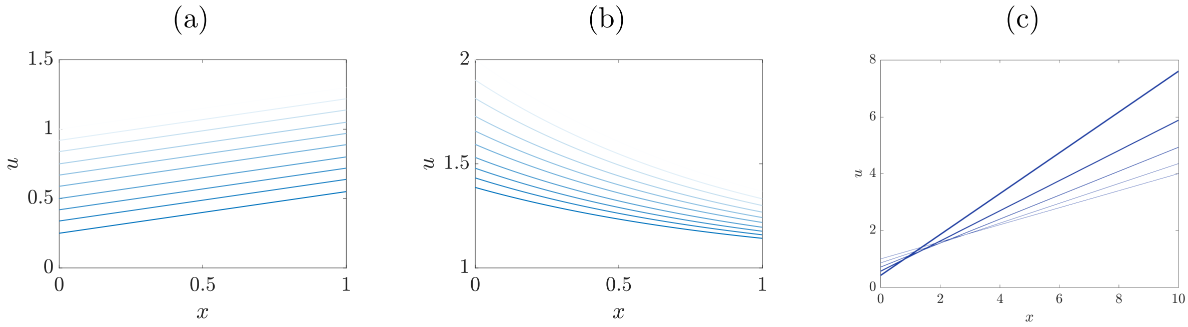

Figure 3: Special solutions that fail (A1). (a) A solution for the advection-diffusion equation (11). (b) Another solution for the advection-diffusion equation (11). (c) A solution for the viscous Burgers' equation (13). Curves with lighter color represents the solution at earlier time.

as the resolution $\Delta x, \Delta t \to 0$. *Here* $\mathbf{1} \in \mathbb{R}^{NM}$ *is the* 1*-vector. Similarly, if* $\mathbf{F}_{p,q}$ *denotes the column of the feature matrix corresponding to the product feature* $\partial_x^p u\, \partial_x^q u$*, then*

$$\left| \int_\Omega \partial_x^p u(x,t) \partial_x^q u(x,t)\, dx\, dt - \mathbf{F}_p^T \mathbf{F}_q \Delta x \Delta t \right| \to 0 \ .$$

Proof. See Appendix B. □

Lemma 4.1 justifies using the normal matrix $\mathbf{F}^T\mathbf{F}\Delta x \Delta t$ to approximate the Gram matrix $\mathbf{G}$, whose (j,k)-entry is the $L^2(\Omega)$ inner product of the j-th and k-th candidatefeature variable, when there are sufficiently many data. Since it is well-known that the Gram matrix is invertible if and only if its associated vectors are linearly independent, we have the following characterization:

Proposition 4.1. *Suppose an evolutionary function* $u : \Omega \to \mathbb{R}$*, i.e.,* $u_t \neq 0$*, satisfies* (1)*. When the number of data is sufficiently large,* $\mathbf{F}_\mathcal{S}^T \mathbf{F}_\mathcal{S}$ *is not invertible if and only if* u *also satisfies a steady-state equation:*

$$\begin{aligned} &0 = F(u, u_x, u_{xx} \ldots; \widetilde{\boldsymbol{\beta}}) := \widetilde{\beta}_0 + \widetilde{\beta}_1 u + \widetilde{\beta}_2 u_x + \widetilde{\beta} u_{xx} + \cdots + \widetilde{\beta}_{p,q} \partial_x^p u \partial_x^q u + \ldots \ , \\ &(x,t) \ in \ \Omega = (0, X_{\max}) \times (0, T_{\max}) \ , \end{aligned} \tag{10}$$

for some $\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^K$ *which has the same support as* $\boldsymbol{\beta}^*$*.*

In this paper, we associate the steady-state equation (10) with the solution-PDE pair $\big(u, \mathcal{F}(\cdot, \boldsymbol{\beta}^*)\big)$. Notice that the steady-state equation is an ordinary differential equation (ODE); hence, verifying (A1) involves solving the steady-state equation then plugging it to the original.

For any PDEs consisting of one feature variable, i.e., $\big|\mathcal{S}(\boldsymbol{\beta}^*)\big| = 1$, with sufficiently many data, the invertibility condition (A1) always holds for any non-steady-state function u, $u_t \neq 0$. This includes non-steady-state solutions of transport equation, heat equation, or inviscid Burgers equation, etc. In the following, we show some non-trivial examples.

Example 4.1. Consider an advection-diffusion equation:

$$u_t = D u_{xx} - v u_x \ , \ 0 < x < X_{\max}, 0 < t < T_{\max} \ , \tag{11}$$

where $D > 0$ is the diffusitivity coefficient, and $v > 0$ is the advection speed. It describes the evolution of the concentration distribution of a non-decaying pollutant in a flowing stream. Suppose we observe a solution u of (11) which also satisfies the steady-state equation :

$$a u_{xx} + b u_x = 0 \tag{12}$$

for some $a, b \in \mathbb{R}$ such that $a^2 + b^2 \neq 0$.

1. If $b=0$, then solving (12) gives $u=A(t)x+B(t)$ for arbitrary functions A and B; plugging this into PDE (11) leads to:

$$\begin{cases} A(t) = A_0 \\ B(t) = -vA_0 t + B_0 \end{cases}, \quad 0 \le t \le T_{\max}$$

   where A_0 and B_0 are arbitrary constants.

2. If $a=0$, (12) gives $u=A(t)$ for an arbitrary function A; plugging this into (11) leads to that u is steady-state solution, hence a contradiction.

3. If $a\neq 0$ and $b\neq 0$, solving (12) gives $u=A(t)+B(t)\exp(-bx/a)$; then (11) implies:

$$\begin{cases} A(t) = A_0 \\ B(t) = B_0 \exp\left((Db^2/a^2 + vb/a)t\right) \end{cases}, \quad 0 \le t \le T_{\max}$$

In summary, with sufficiently many data, if we observe $u(x,t)=A_0x-vA_0t+B_0$ or $u(x,t)=A_0+B_0\exp((Db^2/a^2+vb/a)t-bx/a)$ for arbitrary constants A_0, B_0, $A_0^2+B_0^2\neq 0$, then the invertibility condition (A1) associated with the PDE (11) fails. In Figure 3, we show the special solution $u(x,t)=0.3x-0.75t+1$ in (a) and $u(x,t)=1+\exp(-0.95t-x)$ in (b) for illustration.

Example 4.2. Consider the viscous Burgers' equation:

$$u_t = Du_{xx} + uu_x \tag{13}$$

which is the fundamental model for the dissipative system such as traffic flow. The associated steady-state equation is the following

$$au_{xx} + buu_x = 0 \tag{14}$$

with $a^2+b^2\neq 0$.

1. If $a=0$, then (14) leads to a general solution $u=A(t)$ for an arbitrary function A. However, (13) then forces that u to be a constant; hence, $a\neq 0$.

2. If $b=0$, then (14) leads to a general solution $u=A(t)x+B(t)$ for arbitrary functions A and B. (13) then imposes conditions:

$$\begin{cases} A^2(t) = A'(t) \\ A(t)B(t) - B'(t) = 0 \end{cases} \implies \begin{cases} A(t) = \frac{A_0}{1-A_0t} \\ B(t) = 1 - A_0t \end{cases}$$

   where $A_0\neq 0$ is an arbitrary constant. Since we assume that u is continuous in Ω, $A_0 < 1/T_{\max}$.

3. If $a\neq 0$ and $b\neq 0$, (14) can be transformed to a Riccati equation in terms of u_x, from which we may solve for the general solution of (14) as:

$$u(x,t) = \frac{cA(t)\exp(\frac{bc}{a}x) - cB(t)\exp(-\frac{bc}{a}x)}{A(t)\exp(\frac{bc}{a}x) + B(t)\exp(-\frac{bc}{a}x)} \tag{15}$$

   where c is an arbitrary constant, and A, B are arbitrary functions. By plugging (15) into (13), we immediately see that the common solution exists if and only if $a=2Db$, which is:

$$u(x,t) = \frac{cA_0\exp(cx/(2D)) - cB_0\exp(-cx/(2D))}{A_0\exp(cx/(2D)) + B_0\exp(-cx/(2D))}.$$

   However, this is a steady-state solution, hence a contradiction.

In Figure 3 (c), we show the solution $u(x,t) = 0.3x/(1-0.3t) + 1 - 0.3t$ for illustration.

It is relatively easy to investigate the invertibility condition when the underlying PDE is linear. As for general nonlinear PDEs, various techniques are designed for nonlinear ODEs which can be applied to the steady-state equations. In some cases, such as Fischer's equation and Korteweg-de Vries equation, the solutions to the steady-state equations may involve elliptic integrals, which makes it more difficult to validate the invertibility condition.

By combining Lemma 4.1 with Corollary A.2 and Corollary A.3, we extend the discussion above to the case where noise is included.

Corollary 4.1. *Denoting $\widehat{\mathbf{F}}_k$ as the column of the approximated feature matrix corresponding to the feature $\partial_x^k u$, $k = 0, 1, \cdots, P_{\max}$, then*

$$\left| \int_\Omega \partial_x^k u(x,t)\, dx\, dt - \mathbf{1}^T \widehat{\mathbf{F}}_k \Delta x \Delta t \right| \to |\Omega| \, ||\widehat{\mathbf{F}}_k - \mathbf{F}_k||_\infty$$

as the number of data $N, M \to \infty$. Similarly, if $\widehat{\mathbf{F}}_{k,j}$ denotes the column of the feature matrix corresponding to the product feature $\partial_x^k u\, \partial_x^j u$, then

$$\left| \int_\Omega \partial_x^k u(x,t) \partial_x^j u(x,t)\, dx\, dt - \widehat{\mathbf{F}}_k^T \widehat{\mathbf{F}}_j \Delta x \Delta t \right| \to |\Omega| \, ||\widehat{\mathbf{F}}_k^T \widehat{\mathbf{F}}_j - \mathbf{F}_k^T \mathbf{F}_j||_\infty \, .$$

Proof. By triangle inequality

$$\begin{aligned} \left| \int_\Omega \partial_x^k u(x,t)\, dx\, dt - \mathbf{1}^T \widehat{\mathbf{F}}_k \Delta x \Delta t \right| &\le \left| \int_\Omega \partial_x^k u(x,t)\, dx\, dt - \mathbf{1}^T \mathbf{F}_k \Delta x \Delta t \right| + ||\widehat{\mathbf{F}}_k - \mathbf{F}_k||_1 \Delta x \Delta t \\ &\le o(\Delta x) + o(\Delta t) + |\Omega| \, ||\widehat{\mathbf{F}}_k - \mathbf{F}_k||_\infty \, . \end{aligned}$$

□

Corollary 4.2. *For any $\varepsilon > C'|\Omega| M^{-2/(6+p)}$ with sufficiently large M and N, we have:*

$$\mathbb{P}\Big[\left| \int_\Omega \partial_x^p u(x,t)\, dx\, dt - \mathbf{1}^T \widehat{\mathbf{F}}_p \Delta x \Delta t \right| > \varepsilon \Big] < 2M \exp\Big(-\frac{M^{1/(6+p)} - ||u||_{L^\infty(\Omega)}^2}{2\sigma^2}\Big) \, ,$$

and

$$\mathbb{P}\Big[\left| \int_\Omega \partial_x^p u(x,t) \partial_x^q u(x,t) |\, dx\, dt - \widehat{\mathbf{F}}_p^T \widehat{\mathbf{F}}_q \Delta x \Delta t \right| > \varepsilon \Big] < 8M \exp\Big(-\frac{M^{1/(6+\max\{p,q\})} - ||u||_{L^\infty(\Omega)}^2}{2\sigma^2}\Big) \, .$$

This shows that when there is noise, the result will depend on the accuracy of $||\widehat{\mathbf{F}}_k - \mathbf{F}_k||_\infty$, that a good approximation becomes more important. Using local polynomial approximation, as $||\widehat{\mathbf{F}}_k - \mathbf{F}_k||_\infty \to 0$, the error will reduce accordingly. The size of the domain $|\Omega|$ also effects the result. When the size of the domain increases the error increases, yet, when the size of the domain is fixed, increasing the resolution of the approximation will reduce the error.

4.2 Mutual Incoherence Condition (A2)

Mutual incoherence condition is more challenging to verify. Intuitively, this assumption is closely related to whether the observed solution u displays characteristic movements that are representative to the underlying PDE. Therefore, the data-size, the observed solution of the underlying PDE, and the choice of candidate feature variables have deterministic impacts.

Example 4.3. Diffusion and decaying can cause confusion for PDE identification. As analyzed in the previous subsection, when there is no noise, we may consider the following continuous approximation of (A2) provided that the candidates only include u and u_{xx} and u_{xx} is the only true feature:

$$\frac{|\langle u, u_{xx}\rangle_{L^2(\Omega)}|}{||u_{xx}||^2_{L^2(\Omega)}} \leq 1-\mu. \tag{16}$$

Suppose the observed solution can be expressed as $u(x,t) = A(t)\cos\omega x + B(t)\sin\omega x$ for some $\omega > 0$, then the quantity on the left hand side of (16) is simply $1/\omega^2$. Therefore, when $\omega < 1$, (16) fails; and if $\omega > 1$, it holds for some μ. Due to possibly insufficient sampling, the frequency demonstrated in the data can be lowered, thus (16) may not hold even if ω is large. We may extend the discussion here to u which is analytic along space for every time, and conclude that if the observed solution consists of low frequency components in space, then (A2) fails when u and u_{xx} are the candidates. Furthermore, we mention that the statements above is exactly reversed if u is the true feature instead of u_{xx}.

In Section 5, we look into the plane wave solutions of a transport equation which are completely characterized by their magnitudes and wavenumbers.

4.3 Minimal Eigenvalue Condition (A3)

Minimal eigenvalue condition (A3) is associated with the detectable magnitude of the observed solution of the underlying PDE, rather than specific features determined by the PDE. For instance, if the underlying PDE consists of only one feature variable, such as the cases in transport equation and heat equation, then (A3) is simply requiring that, with sufficiently many data, the $L^2(\Omega)$-norm of the feature variable exceeds $C_{\min} > 0$. We present an example to illustrate this idea.

Example 4.4. If the underlying PDE is advection-diffusion equation, then the associated normal matrix $\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}}$ is:

$$\begin{bmatrix} ||u_x||^2_{L^2(\Omega)} & \langle u_x, u_{xx}\rangle_{L^2(\Omega)} \\ \langle u_x, u_{xx}\rangle_{L^2(\Omega)} & ||u_{xx}||^2_{L^2(\Omega)} \end{bmatrix}$$

whose minimal eigenvalue $\Lambda_{\min}$ can be proved to satisfy:

$$\begin{aligned} 0 \leq \Lambda_{\min} \leq \min\{||u_x||^2_{L^2(\Omega)}, ||u_{xx}||^2_{L^2(\Omega)}\}\ , &\quad \text{If } \langle u_x, u_{xx}\rangle_{L^2(\Omega)} \geq 0\ ; \\ \min\{||u_x||^2_{L^2(\Omega)}, ||u_{xx}||^2_{L^2(\Omega)}\} \leq \Lambda_{\min}\ , &\quad \text{If } \langle u_x, u_{xx}\rangle_{L^2(\Omega)} \leq 0\ . \end{aligned}$$

In both cases, (A3) holds more easily if the magnitudes of the underlying feature variables are large. An interesting aspect of this example is that, when u_x and u_{xx} are negatively correlated, (A3) immediately holds by setting $C_{\min} = \min\{||u_x||^2_{L^2(\Omega)}, ||u_{xx}||^2_{L^2(\Omega)}\}$.

5 Numerical Experiments

We present numerical results of the proposed model. First, we present results for various PDEs and their identifications. Secondly, we explore the connections between mutual incoherence condition (A2) vs. data size and the quality of observed data, then numerically validate Theorem 3.1 by exploring the bound of λ.

We use Neumann boundary condition in this paper, i.e., the first-order partial derivatives are set to 0. For local polynomial fitting, we symmetrically pad the time and space domain. The implementation becomes simpler with this setting, and the computation is 90% faster. More advanced techniques for addressing the boundary estimations can be found in [23, 24].

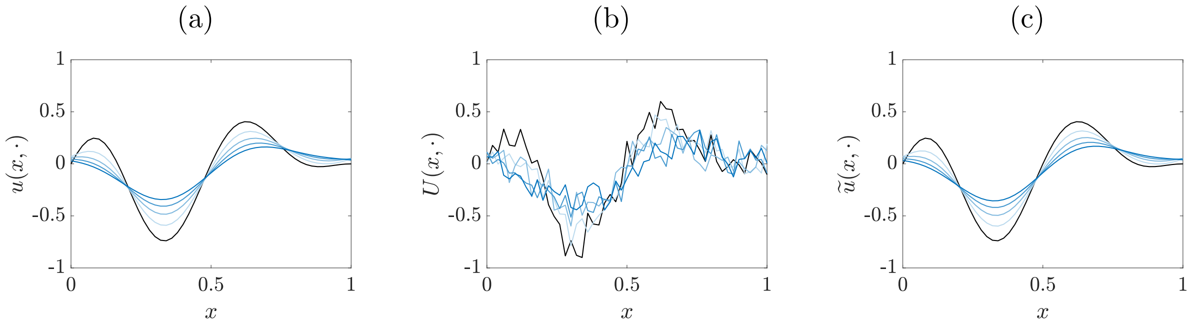

Figure 4: Heat equation (17). (a) Numerical solution of $u_t = 0.1u_{xx}$ with initial condition: $f(x) = 5(1-x)^2 x\cos(3\pi x),\ 0 \le x \le 1$. (b) Data obtained from (a) with additive Gaussian noise of intensity $\sigma = 0.1$. (c) Numerical solution of the identified PDE: $u_t = 0.0939u_{xx}$.

5.1 Identification of PDEs using ℓ_1-PsL: Examples

We apply the proposed ℓ_1-PsL method to identify classical PDEs under noisy observation. We consider the heat equation (second-order linear equation) and Burgers' equation (first-order non-linear equation). Each of them plays a fundamental role in modeling physical phenomenon and demonstrates characteristic behaviors shared by more complex systems, such as dissipation and shock-formation. We refer the readers to [25] for derivations and more.

Heat Equation is one of the most fundamental PDEs in physics as well as mathematics. It models the variation of the temperature distribution along a conductive material as the time proceeds. For a metal ring, one can equip the heat equation with a periodic boundary condition:

$$\begin{aligned} &u_t(x,t) = \nu u_{xx}(x,t)\ , 0 < x < 1,\ 0 < t < T_{\max}\ . \\ &u(x,0) = f(x)\ , 0 \le x \le 1\ ; \\ &u(0,t) = u(1,t)\ , 0 \le t \le T_{\max} \end{aligned} \tag{17}$$

Here $u(x,t)$ denotes the temperature of the bar at location $0 \le x \le 1$ and time $0 \le t \le 1$, $\nu > 0$ is the material conductivity, and $f(x)$ describes the initial heat distribution on the ring.

Figure 4 (a) shows the numerical solution when the initial condition is $f(x) = 5(1-x)^2 x\cos(3\pi x)$, $0 \le x \le 1$, terminal time is $T_{\max} = 0.1$, and the conductivity constant is $\nu = 0.1$. We solve (17) by the forward time centered space scheme using $\Delta t = 6.6667 \times 10^{-7} (= 0.0033/5000)$ and $\Delta x = 0.02$, then we downsample the solution by 5000 in time, resulting in $N = 30$ and $M = 50$. (b) demonstrates the noisy data ($\sigma = 0.1$), from which, we apply the ℓ_1-PsL and identify the correct PDE which reproduce the underlying dynamic in (c). Notice the recovered coefficient 0.0939 for the true value of 0.1.

Figure 5 illustrates the dependence of the feature variable identification on the choice of λ, under different levels of noise: (a) $\sigma = 0.01$, (b) $\sigma = 0.1$, and (c) $\sigma = 0.3$. The dashed curves represent the coefficients of the false features, and the red one represents that of the correct one. As λ increases, the false feature variables are filtered out, while the correct one u_{xx} remains active. As the noise level increases, it becomes more difficult to separate the correct feature variables from the false: The vanishing points of the dashed curves are approaching to that of the red curve.

Burgers' Equation is a simplified Navier-Stokes equation when the pressure gradient is zero. The one-dimensional case equipped with 0-Dirichlet boundary condition is:

$$\begin{aligned} &u_t(x,t) = -(\frac{1}{2}u^2)_x + \nu u_{xx}(x,t)\ , 0 < x < 1,\ 0 < t < T_{\max}\ . \\ &u(x,0) = f(x)\ , 0 \le x \le 1\ ; \\ &u(0,t) = u(1,t) = 0\ , 0 \le t \le T_{\max} \end{aligned} \tag{18}$$

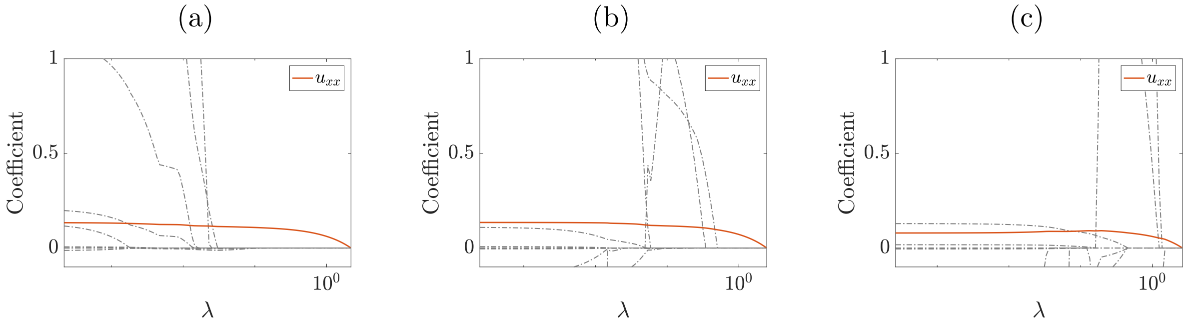

Figure 5: Heat equation (17). Plots of the coefficients of the candidate feature variables for various parameters λ in the ℓ_1-PsL model under different noise levels: (a) $\sigma = 0.01$, (b) $\sigma = 0.1$, $\sigma = 0.3$. The dashed lines correspond to coefficients of the false feature variables, while the correct one is highlighted by red. As λ increases, only the coefficient of the correct feature variable remain non-zero.

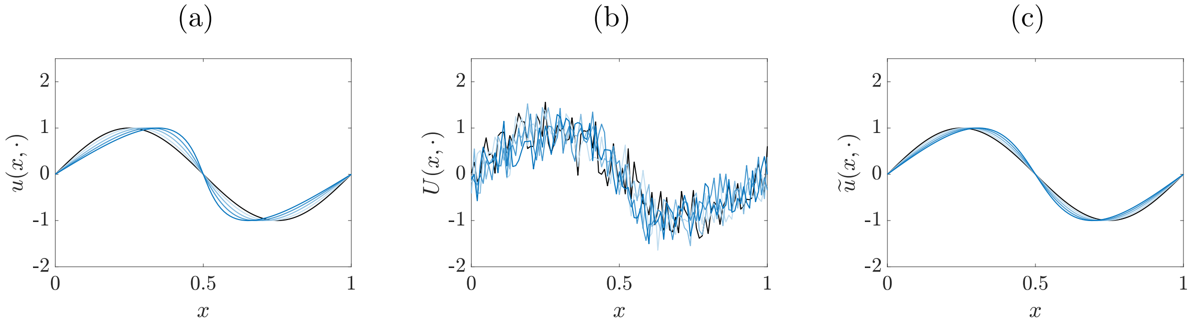

Figure 6: Inviscid Burgers' equation (18). (a) Numerical solution of the inviscid Burgers' equation: $u_t = -uu_{xx}$ with initial condition: $f(x) = \sin(2\pi x)$, $0 \leq x \leq 1$. (b) Data obtained from (a) with additive Gaussian noise of intensity $\sigma = 0.3$. (c) Numerical solution of the identified PDE: $u_t = -0.5946uu_x$.

When the viscosity coefficient $\nu = 0$, the PDE (18) is often called inviscid Burgers' equation, and it is closely related to the conservation law. If $\nu > 0$, then (18) is referred to as viscous Burgers' equation, which takes the energy dissipation into consideration.

In Figure 6 (a), we show the numerical solution of the inviscid Burgers' equation with $f(x) = \sin(2\pi x)$, $0 \leq x \leq 1$, $T_{\text{max}} = 0.1$, using $\Delta x = 0.01$ and $\Delta t = 3.5714 \times 10^{-7} (= 0.0018/5000)$. We downsample the solution in times by 5000, which results in $N = 56$ and $M = 100$. (b) shows the noisy data ($\sigma = 0.3$) and (c) displays the numerical solution of the PDE identified by ℓ_1-PsL. The coefficient of the identified feature variable -0.5946 is smaller than the true value -1; this is due to insufficient time resolution (See Table 1) and the smoothing effects of the local polynomial regression technique. Table 1 shows the coefficient accuracy as the time resolution increases. The coefficient value approaches the true value of -1, as time resolution gets higher. Figure 7 shows the dependence of the coefficients of the candidate feature variables against the choice of λ. As the noise increases, the coefficient of the correct feature variable becomes less accurate, and those of the false feature variables becomes more significant.

Figure 8 (a) shows the numerical solution of the viscous Burgers' equation with $f(x) = \sin^2(4\pi x) + \sin^3(2\pi x)$, $0 \leq x \leq 1$, $T_{\text{max}} = 0.1$, and $\nu = 0.01$, while keeping the grid same as above. (b) shows the noisy data ($\sigma = 0.5$) and (c) displays the solution of the PDE identified by ℓ_1-PsL. Higher level of noise submerges the shock formation determined by the feature uu_x, and the smoothing effects of the local polynomial regression analogous to the diffusion term u_{xx} becomes obvious. This is clear in Figure 9. After all the false feature variables have vanished, when the noise increases, ℓ_1-PsL filters out the feature u_{xx} later than uu_x.

Time resolution ($N_0 = 100^{7/8}$)	N_0	$10N_0$	$100N_0$
Estimated coefficient for uu_x	-0.6794	-0.7499	-0.8005

Table 1: Inviscid Burgers' equation (18). Increased time resolution results in higher accuracy of the estimated coefficient of the identified feature variable: uu_x in the inviscid Burgers' equation. The true value is -1, and the space resolution is remained at $M = 100$.

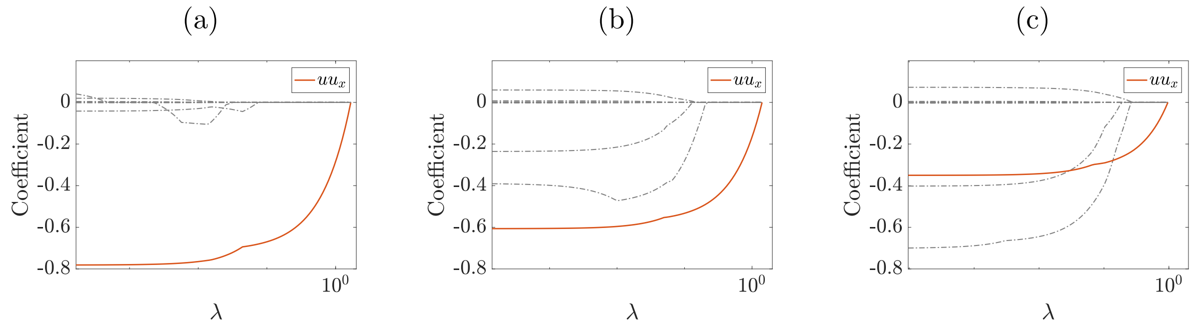

Figure 7: Inviscid Burgers' equation (18). Plots of the coefficients of the candidate feature variables for various parameters λ in the ℓ_1-PsL model under different noise levels: (a) $\sigma = 0.1$, (b) $\sigma = 0.5$, $\sigma = 1.0$. The dashed lines correspond to coefficients of the false feature variables, while the correct one is highlighted by red.

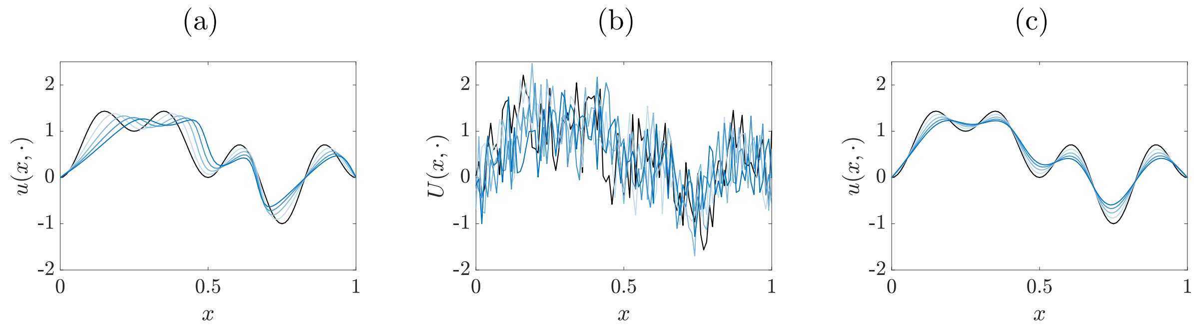

Figure 8: Viscous Burgers' equation (18). (a) Numerical solution of the viscous Burgers' equation: $u_t = -uu_x + 0.01u_{xx}$ with initial condition: $f(x) = \sin^2(4\pi x) + \sin^3(2\pi x);,\ 0 \leq x \leq 1$. (b) Data obtained from (a) with additive Gaussian noise of intensity $\sigma = 0.5$. (c) Numerical solution of the identified PDE: $u_t = -0.2115uu_x + 0.0131u_{xx}$.

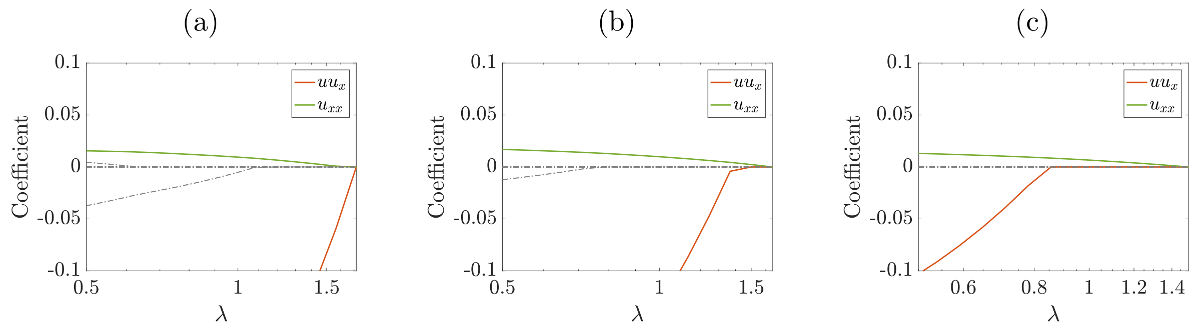

Figure 9: Viscous Burgers' equation (18). Plots of the coefficients of the candidate feature variables for various parameters λ in the ℓ_1-PsL model under different noise levels: (a) $\sigma = 0.1$, (b) $\sigma = 0.5$, and (c) $\sigma = 1.0$. We focus on the range of λ where the false feature variables have vanished. The dashed lines correspond to coefficients of the false feature variables, while the correct ones are highlighted by red and green.

5.2 Mutual Incoherence Condition and Data Size

The deterministic version of (A2) is commonly employed as an important sufficient condition for exact sparse recovery. In Section 4.2, we relate this condition to the exhibition of characteristic behaviors of the underlying PDE, such as advection (in relation to u_x) and diffusion (in relation to u_{xx}). It is known that justifying the deterministic mutual incoherence property is NP-hard [26], thus validating (A2) for an unknown PDE and an arbitrary set of candidate feature variables is impractical. Instead, we focus on a concrete example where (A2) holds with high probability provided as the data size increases.

Let's consider the transport equation:

$$\begin{aligned} u_t(x,t) &= a u_x \ , x \in \mathbb{R}, \ 0 < t < T_{\max} \ . \\ u(x,0) &= f(x) \ , x \in \mathbb{R} \ . \end{aligned} \tag{19}$$

where a represents the advection speed. Notice that, although (19) specifies the PDE satisfied by u over $\mathbb{R}$, the data $\mathcal{D}$ is only restricted within $0 \leq x \leq X_{\max}$. The solution to (19) is simply $u(x,t) = f(x+at)$, and our discussion below is based on the choices: $a = -2$, $X_{\max} = 1$, $T_{\max} = 1$, and $f(x) = 2\sin 4x$, so that $u(x,t) = 2\sin(4x - 8t)$.

Figure 10 shows the ℓ_∞-norm $||\widehat{\mathbf{F}}_{\mathcal{S}^c}^T \widehat{\mathbf{F}}_{\mathcal{S}} \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}||_\infty$ for noise level $\sigma = 0.1$ in (a) and $\sigma = 0.3$ in (b) as the data size increases. If this ℓ_∞-norm is below 1, (A2) holds. In both cases, (A2) is satisfied with high probability when the size of data is large. As the noise level increases, the number of data required for satisfying (A2) also rises. When the data size is relatively small, adding some noise to the data may lead to satisfying (A2), even if it does not hold when the data is clean. Different from reducing $||\widehat{\mathbf{F}}_{\mathcal{S}^c}^T \widehat{\mathbf{F}}_{\mathcal{S}} \big(\widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}||_\infty$ by enhancing the resolution, the ℓ_∞-norm reduction induced by adding noise is unstable, and the underlying dynamics can be deformed.

5.3 Mutual Incoherence Condition and the Observed Solution

Let's consider a transport equation $u_t = -(c/\omega)u_x$, $0 \leq x \leq X_{\max}$, $0 \leq t \leq T_{\max}$ which admits the solutuon: $u(x,t) = A\sin(\omega x - ct)$ for $A \neq 0$. For PDE identification, we choose a constant term, u, u_x and u_{xx} to be the candidate feature variables. Provided with sufficiently many data, we note that both (A1) and (A3) hold trivially for transport equation, and the deterministic version of (A2) holds if and only if

$$H(\omega, c) := \frac{\max\Big\{\Big|\langle 1, u_x\rangle_{L^2(\Omega)}\Big|, \Big|\langle u, u_x\rangle_{L^2(\Omega)}\Big|, \Big|\langle u_{xx}, u_x\rangle_{L^2(\Omega)}\Big|\Big\}}{||u_x||_{L^2(\Omega)}^2} < 1 - \mu. \tag{20}$$

In this case, if (20) holds, all the assumptions in Section 3.1 are satisfied; and we can apply Theorem 3.1 to guarantee recovery of u_x when λ is greater than a threshold.

We plot the contour of $H(\omega, c)$ in Figure 11 (a) when $X_{\max} = T_{\max} = A = 0.5$. The yellow region represents $H(\omega, c) > 1$, thus (A2) fails; the light green region is when $0.5 < H(\omega, c) < 1$, and the dark green region is when $H(\omega, c) < 0.5$. Although (A2) is only a sufficient condition for recovering the feature, we identify a wrong feature variable u as λ increases if the pair (ω, c) is deep inside the yellow region ($H(\omega, c) \gg 1$), see (b) and (c). When (ω, c) resides inside the dark green region ($H(\omega, c) < 0.5$), the correct feature u_x is chosen, as shown in (d).

As we vary the domain size $X_{\max}$, the observation time $T_{\max}$, or the magnitude of the the solution A, the contour of $H(\omega, c)$ modifies accordingly. In Figure 12, we fix $T_{\max} = 0.5$ and $A = 0.1$; as we increase $T_{\max}$, the yellow region gradually shrinks its area and flattens toward the c-axis. In Figure 13, we fix $X_{\max} = 0.5$ and $A = 0.1$; as the $X_{\max}$ gets larger, the yellow region also shrinks its area and flattens toward the ω-axis. The effects of increasing the magnitude A are illustrated in Figure 14, where the yellow region shrinks along the diagonal direction towards

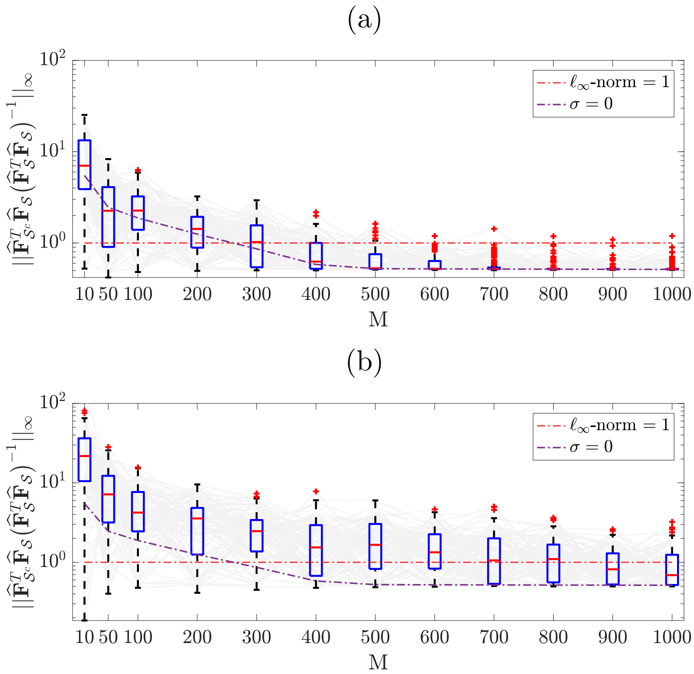

Figure 10: Plot of $||\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}}\big)^{-1}||_\infty$ computed based on the noisy data from $u(x,t) = 2\sin(4x - 8t)$ satisfying the PDE (19) with Gaussian noise of intensity (a) $\sigma = 0.1$, (b) $\sigma = 0.3$. The boxplots are obtained by 100 independent experiments. As the data size increases ($M = N^{8/7} \to \infty$), this ℓ_∞-norm reduces below 1 on average, indicating that the mutual incoherence condition (A2) holds with high probability with sufficiently many data in this case.

the origin. These shrinking behaviors in Figure 12, 13 and 14 are expected. When the shape of $u(\cdot,t)$ for any t contains sufficient variations (larger ω), increasing the domain of observation, $X_{\max}$, will introduce more complexities into the data that help to distinguish different feature variables. Therefore, the yellow region shrinks and flattens towards the c-axis. Similarly, the yellow region shrinks and flattens, as we increase the observation time $T_{\max}$. Changing the magnitude A magnifies any variations in both space and time, hence the yellow region recedes towards the origin.

5.4 Numerical Validation of Theorem 3.1: λ Bound

We numerically demonstrate the implication of Theorem 3.1. In particular, we show that the lower bound of the ℓ_1-regularization parameter λ when $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$ holds, behaves as $\sim \sqrt{\ln N/N^{4/7}}$ as the data size $N \to \infty$ (recall that $M = N^{8/7}$). Taking $u(x,t) = 2\sin(3x - 8t)$ as a solution of the transport equation $u_t = -(8/3)u_x$ and setting $X_{\max} = 2$, $T_{\max} = 1$, Figure 15 shows the lower bound of λ that properly recovers the feature variables together with the curve which is some multiple of $\sqrt{\ln N/N^{4/7}}$. The multiple is computed using simple linear regression. The data for (a) is contaminated by Gaussian noise with $\sigma = 0.1$ and for (b), the noise level is $\sigma = 0.3$. In both cases, the variation of the lower bound for λ is well captured by the ratio $\sqrt{\ln N/N^{4/7}}$ as $N \to \infty$.

6 Conclusion

A formal statistical analysis on the identification of PDE are relatively new in the statistical literature. In this article, we assume that the differential equation governing the dynamic system can be represented as a liner combination of various linear and nonlinear differential terms.

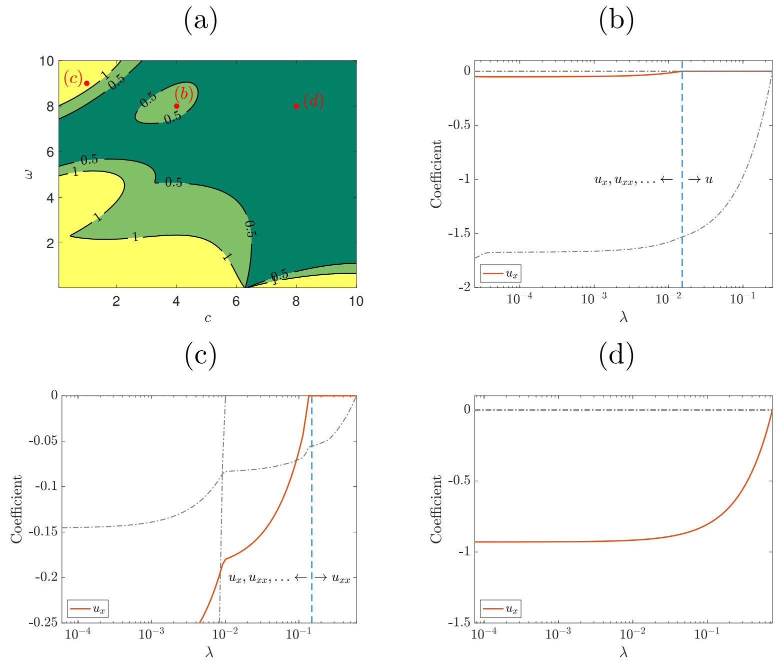

Figure 11: Example of a transport equation $u_t = -(c/\omega)u_x$, and $A = X_{\max} = T_{\max} = 0.5$. (a) Contour plot of $H(\omega, c)$ defined in (20). The yellow region is when $H(\omega, c) > 1$, the light green region when $0.5 < H(\omega, c) < 1$, and the dark green region when $H(\omega, c) < 0.5$. (b) Wrong feature is identified when $\omega = 8, c = 4$. (c) Wrong feature is identified when $\omega = 9, c = 1$. Both points in (b) and (c) belong to the yellow region where (A2) fails. (d) Correct feature is identified when $\omega = 8, c = 8$, which lies in the dark green region where (A2) holds.

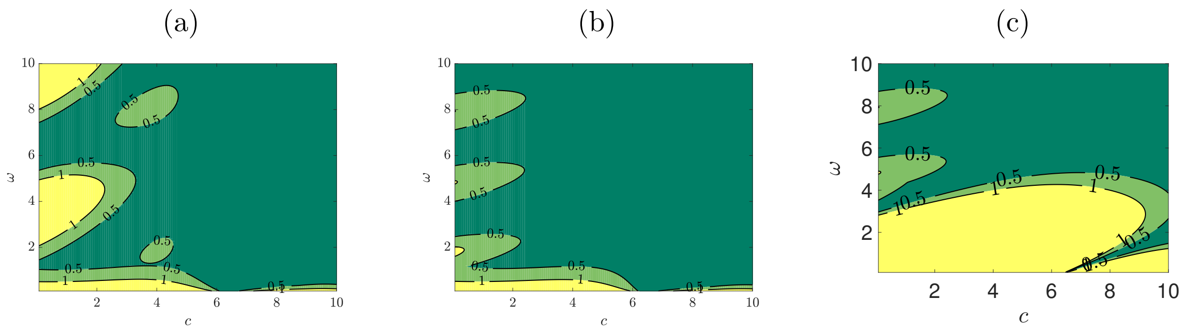

Figure 12: Effects of varying domain of observation $X_{\max}$ on the contour of $H(\omega, c)$. (a) $X_{\max} = 0.1$; (b) $X_{\max} = 0.5$; (c) $X_{\max} = 1$. In all cases, $T_{\max} = 0.5$ and $A = 0.1$. As $X_{\max}$ increases, the yellow region ($H(\omega, c) > 1$) shrinks and flattens towards the c-axis.

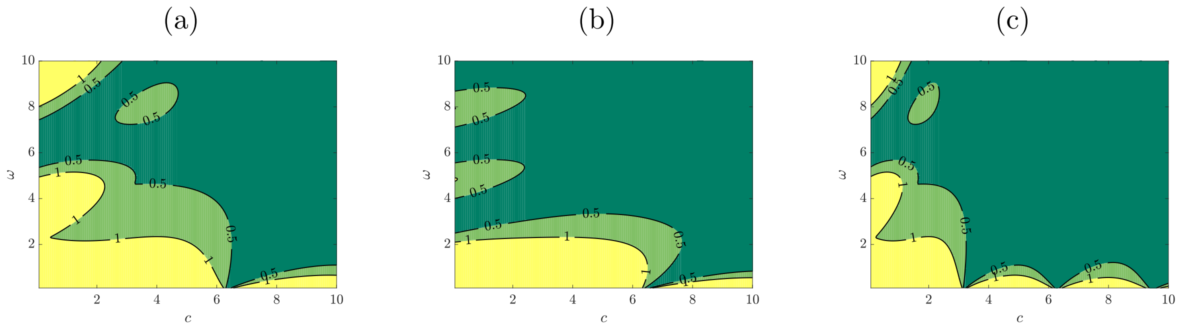

Figure 13: Effects of varying time of observation $T_{\max}$ on the contour of $H(\omega, c)$. (a) $T_{\max} = 0.1$; (b) $T_{\max} = 0.4$; (c) $T_{\max} = 0.8$. In all cases, $X_{\max} = 0.5$ and $A = 0.1$. As $T_{\max}$ increases, the yellow region ($H(\omega, c) > 1$) shrinks and flattens towards the ω-axis.

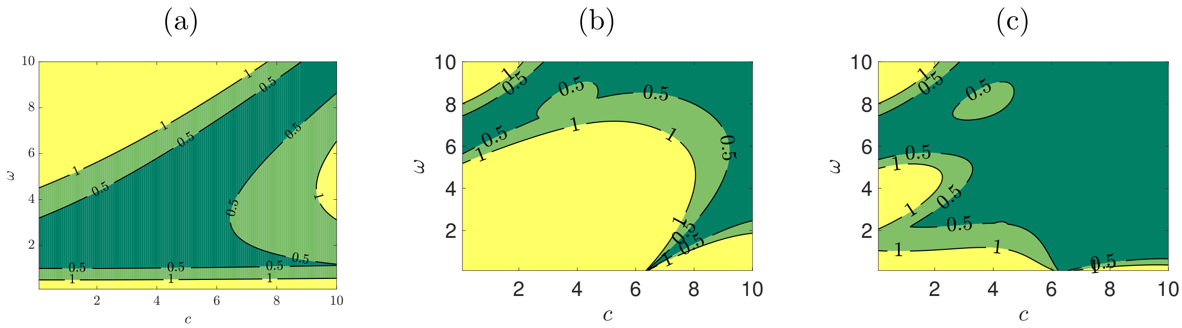

Figure 14: Effects of varying solution magnitude A on the contour of $H(\omega, c)$. (a) $A = 0.01$; (b) $A = 0.3$; (c) $A = 1$. In all cases, $X_{\max} = 0.5$ and $T_{\max} = 0.5$. As A increases, the yellow region ($H(\omega, c) > 1$) shrinks along the diagonal direction towards the origin.

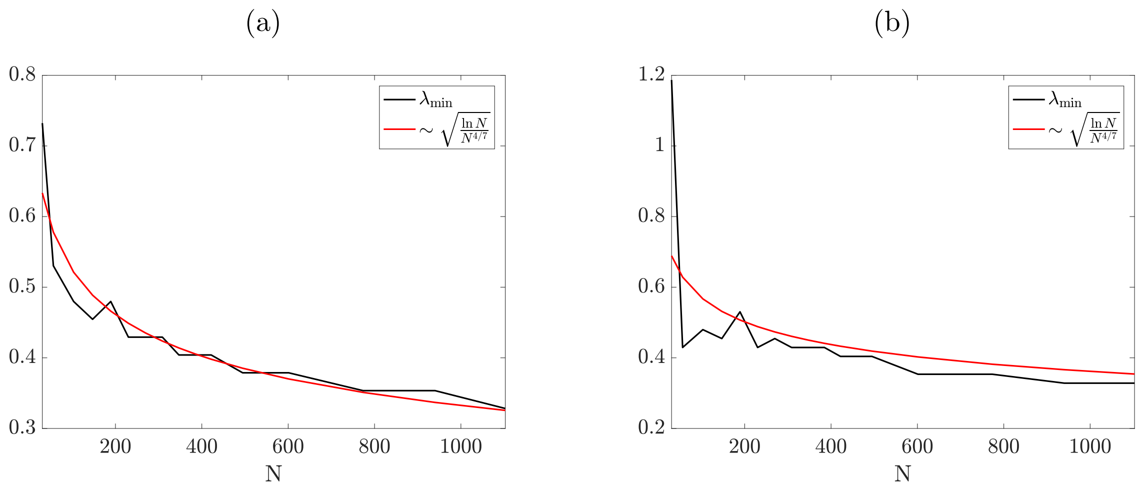

Figure 15: As predicted in Theorem 3.1, the minimal λ required to properly recover $\mathcal{S}(\boldsymbol{\beta}^*)$ decreases in the same order as $\sqrt{\ln N / N^{4/7}}$ when $N \to \infty$. The underlying PDE is $u_t = -(8/3)u_x$, and the data comes from a solution $2\sin(3x - 8t)$ with $X_{\max} = 2$, $T_{\max} = 1$, and noise level $\sigma = 0.1$ for (a) and $\sigma = 0.3$ for (b). The multipliers in front of $\sqrt{\ln N / N^{4/7}}$ are found via simple linear regression.

We employ local polynomial fitting and apply ℓ_1-PsL for model selection to identify correct differential terms. With a properly designed bandwidth parameter, local polynomial regression is used to construct feature matrix. An exponential decay rate for achieving strict dual feasibility is obtained, when the constructed feature matrix satisfies mutual incoherence property. We provide intuitive insights on how the minimum eigenvalue condition and mutual incoherence property can be understood in the PDE context. We propose our models in the framework of measurement error models to avoid some critical problems of the existing method that includes 1) convergence problem of the Least square methods, 2) high computational cost due to iteratively solving PDE numerically in the estimation proceure; and 3) high computational cost due to complicated optimization problem.

There is also a cost associated with the proposed method both from the theoretical and practical points of view. F irstly, a detailed investigation of our proof unveils that the conditions (A2) and (A3) are essential ingredients for obtaining support recovery results in ℓ_1-PsL. Since entries of the estimated feature matrix via local smoothing method are involved with stochasticity, it is desirable to obtain certain probability bounds for the events to happen. It turns out that getting the tight bound is a difficult problem, since we do not know the exact form of the distribution on each entry of gram matrix. Also, as noted by Liang and Wu [1], another caveat of our method is that it needs a fair amount of observations for good estimation results for both the response and predictor variables. Particularly, it is a well-known that the local kernel smoothing requires a relatively large sample size for the accurate estimates of state variables and their derivatives. A requirement of a large observations for the correct support recovery is also demonstrated by the experiments in Section 4 and 5. We are currently investigating how to combine our proposed ℓ_1-PsL method with other existing methods to overcome the computational problems of the existing methods, while at the same time put our efforts to see if we can obtain exponentially decaying tail bounds for mutually incoherence quantity for the estimated design matrix, given the ground truth design matrix satisfies mutual incoherence condition.

A Proof of Theorem 3.1

By the KKT-condition, any minimizer $\boldsymbol{\beta}$ of (3) satisfies:

$$-\frac{1}{NM}\widehat{\mathbf{F}}^T(\widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta}) + \lambda\mathbf{z} = 0\ , \text{ for } \mathbf{z} \in \partial||\boldsymbol{\beta}||_1\ , \tag{21}$$

where $\partial||\boldsymbol{\beta}||_1$ denotes the subdifferential of $||\boldsymbol{\beta}||_1$. Let $\Delta\mathbf{u}_t = \widehat{\mathbf{u}}_t - \mathbf{u}_t$, $\Delta\mathbf{F} = \widehat{\mathbf{F}} - \mathbf{F}$ denote the error terms, and notice that $\widehat{\mathbf{u}}_t = \widehat{\mathbf{F}}\boldsymbol{\beta}^* - \Delta\mathbf{F}\boldsymbol{\beta}^* + \Delta\mathbf{u}_t$, thus from (21), we get

$$\widehat{\mathbf{F}}^T\widehat{\mathbf{F}}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + \widehat{\mathbf{F}}^T(\Delta\mathbf{F}\boldsymbol{\beta}^* - \Delta\mathbf{u}_t) + \lambda NM\mathbf{z} = 0\ \ . \tag{22}$$

We decompose (22) as follows:

$$\begin{bmatrix} \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}} & \widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}^c} \\ \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}} & \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}^c} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} - \boldsymbol{\beta}^* \\ \boldsymbol{\beta}_{\mathcal{S}^c} \end{bmatrix} + \begin{bmatrix} \widehat{\mathbf{F}}_{\mathcal{S}}^T \\ \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \end{bmatrix} (\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* - \Delta\mathbf{u}_t) + \lambda NM \begin{bmatrix} \mathbf{z}_{\mathcal{S}} \\ \mathbf{z}_{\mathcal{S}^c} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}\ . \tag{23}$$

Suppose $(\check{\boldsymbol{\beta}}, \check{\mathbf{z}}) \in \mathbb{R}^K \times \mathbb{R}^K$ is a pair obtained by the primal-dual witness construction, where $\check{\boldsymbol{\beta}}_{\mathcal{S}^c} = 0$ and $\check{\mathbf{z}}$ is an element of the subdifferential of $||\check{\boldsymbol{\beta}}||_1$. Plugging $(\check{\boldsymbol{\beta}}, \check{\mathbf{z}})$ into (23) gives:

$$\check{\mathbf{z}}_{\mathcal{S}^c} = \widehat{\mathbf{F}}_{\mathcal{S}^c}^T\widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}})^{-1}\mathbf{z}_{\mathcal{S}} - \frac{1}{\lambda MN}\widehat{\mathbf{F}}_{\mathcal{S}^c}^T\boldsymbol{\Pi}(\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* - \Delta\mathbf{u}_t)\ , \tag{24}$$

where $\boldsymbol{\Pi} = \mathbf{I} - \widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}_{\mathcal{S}}^T\widehat{\mathbf{F}}_{\mathcal{S}})^{-1}\widehat{\mathbf{F}}_{\mathcal{S}}^T$ is an orthogonal projection. By the complementary slackness condition, $|\check{\mathbf{z}}_j| < 1$ implies $\check{\boldsymbol{\beta}}_j = \mathbf{0}$, which guarantees the proper support recovery. By (A1), we can focus on proving that, as $N, M \to \infty$, $\mathbb{P}\big[\max_{j\in\mathcal{S}^c}|\widetilde{Z}_j| \geq s\big] \to 0$, for $\widetilde{Z}_j = [\widehat{\mathbf{F}}_{\mathcal{S}^c}]_j^T \mathbf{\Pi}\frac{\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}^*_{\mathcal{S}} - \Delta\mathbf{u}_t}{\lambda NM}$, $[\widehat{\mathbf{F}}_{\mathcal{S}^c}]_j$ is the j-th column of $\widehat{\mathbf{F}}_{\mathcal{S}^c}$. By the following lemma, we claim that to prove Theorem 3.1, it suffices to bound ℓ_∞-norm of the PDE estimation error $\boldsymbol{\tau}$.

Lemma A.1. *For any $\varepsilon > 0$:*

$$\mathbb{P}\big[\max_{j\in\mathcal{S}^c}|\widetilde{Z}_j| > \varepsilon\big] \leq \mathbb{P}\Big[||\boldsymbol{\tau}||_\infty > \frac{\lambda\varepsilon}{\sqrt{K}}\Big] \,. \tag{25}$$

Proof.

$$\begin{aligned}\mathbb{P}\Big[||\widehat{\mathbf{F}}^T_{\mathcal{S}^c}\mathbf{\Pi}\frac{\boldsymbol{\tau}}{\lambda NM}||_\infty > \varepsilon\Big] &\leq \mathbb{P}\Big[||\widehat{\mathbf{F}}^T\mathbf{\Pi}\frac{\boldsymbol{\tau}}{\lambda NM}||_2 > \varepsilon\Big] \leq \mathbb{P}\Big[||\widehat{\mathbf{F}}||_2\,||\frac{\boldsymbol{\tau}}{\lambda NM}||_2 > \varepsilon\Big] \\ &\leq \mathbb{P}\Big[||\boldsymbol{\tau}||_2 > \lambda\varepsilon\sqrt{\frac{NM}{K}}\Big] \leq \mathbb{P}\Big[||\boldsymbol{\tau}||_\infty > \frac{\lambda\varepsilon}{\sqrt{K}}\Big]\end{aligned}$$

□

A.1 Sufficient conditions for bounding $\widehat{\mathbf{u}}_t - \mathbf{u}_t$

Proposition A.1. *For any fixed $i = 0, \cdots, M-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist finite positive constants $A(X_i), \overline{C}(X_i), C, L, Q$ which do not depend on the temporal sample size N, such that for any $\gamma > 0$ and ε^*_N satisfying:*

$$\varepsilon^*_N(X_i) >$$
$$\max\{2|C^*(X_i)|h_N^2, \frac{12\mathcal{K}^*_{\max}B_N}{Nh_N^2}, 12A(X_i)B_N^{1-s}, \frac{12B_N(C\log N + \gamma)\log N}{h_N^2 N}, 12\overline{C}(X_i)\sqrt{\frac{2\ln 1/h_N}{h_N^3 N}}\}\,,$$

where B_N is an arbitrary increasing sequence $B_N \to \infty$ as $N \to \infty$, we have:

$$\mathbb{P}\Big[\sup_{t\in[0,T]}|\Delta u_t(X_i,t)| > \varepsilon^*_N(X_i)\Big] < 2N\exp(-\frac{C_N^2}{2\sigma^2}) + Qe^{-L\gamma} + 4\sqrt{2}\eta^4\exp(-72N)\,, \tag{26}$$

where $C_N = B_N - ||u||_{L^\infty(\Omega)}$.

Proof. In the following argument, we fix $i = 0, \cdots, M-1$ and omit the dependence of the constants on X_i in the notations. Let B_N be a sequence of increasing positive numbers such that $B_N \to \infty$ as $N \to \infty$, then define the truncated estimate:

$$\begin{aligned}\widehat{u_t}^B(X_i,t) &= \frac{1}{Nh_N^2}\sum_{n=0}^{N-1}\mathcal{K}^*(\frac{t_n - t}{h_N})U_i^n I\{|U_i^n| < B_N\} \\ &= \frac{1}{h_N^2}\iint_{|y|<B_N}\mathcal{K}^*(\frac{z-t}{h_N})y\, df_N(z,y)\end{aligned}$$

where $f_N(\cdot,\cdot) := f_N(\cdot,\cdot|X_i)$ is the empirical distribution of (t_n, U_i^n). Construct another increasing sequence by $C_N = B_N - ||u||_{L^\infty(\Omega)}$ (which is positive with a sufficiently large N), and notice that for any $\varepsilon \geq \frac{\mathcal{K}^*_{\max}B_N}{Nh_N^2}$:

$$\begin{aligned}&\mathbb{P}\Big[\sup_t|\widehat{u_t}(X_i,t) - \widehat{u_t}^B(X_i,t)| > \varepsilon\Big] = \mathbb{P}\Big[\sup_t|\frac{1}{Nh_N^2}\sum_{n=0}^{N-1}\mathcal{K}^*(\frac{t_n-t}{h_N})U_i^n I\{|U_i^n| \geq B_N\}| > \varepsilon\Big] \\ &\leq \mathbb{P}\Big[\frac{\mathcal{K}^*_{\max}}{Nh_N^2}\sum_{n=0}^{N-1}|U_i^n|I\{|U_i^n| \geq B_N\} > \varepsilon\Big] \leq \mathbb{P}\Big[\exists n = 0, 1, \cdots, N-1,\ |U_i^n| \geq B_N\Big] \\ &= \mathbb{P}\Big[\max_{n=0,1,\cdots,N-1}|U_i^n| \geq B_N\Big] \leq \mathbb{P}\Big[\max_{n=0,1,\cdots,N-1}|U_i^n - u_i^n| \geq C_N\Big] \leq 2N\exp(-\frac{C_N^2}{2\sigma^2})\,.\end{aligned}$$

On the other hand, from Proposition 1 of [6]:

$$E|\widehat{u}_t(X_i,t)-\widehat{u}_t^{\,B}(X_i,t)|\le AB_N^{1-s}\ .$$

for $A=\int|\mathcal{K}(\zeta)|\,d\zeta\times\sup_t\int|y|^s f(t,y|X_i)\,dy$ with $f(\cdot,\cdot|X_i)=:f(\cdot,\cdot)$ as the distribution of $(t,U(X_i,t))$; hence for any $\varepsilon_{1,N}\ge\max\{\frac{2\mathcal{K}^*_{\max}B_N}{Nh_N^2},2AB_N^{1-s}\}$, we have:

$$\mathbb{P}\Big[\sup_t|\widehat{u}_t(X_i,t)-\widehat{u}_t^{\,B}(X_i,t)-(E(\widehat{u}_t(X_i,t)-\widehat{u}_t^{\,B}(X_i,t)))|>\varepsilon_{1,N}\Big]\le 2N\exp(-\frac{C_N^2}{2\sigma^2})\ . \quad (27)$$

Following [6], we decompose the truncated estimate as follows:

$$\widehat{u}_t^{\,B}(X_i,t)=E\Big[\widehat{u}_t^{\,B}(X_i,t)\Big]+\frac{1}{\sqrt{N}}\rho_N(X_i,t)+e_N(X_i,t)\ ,$$

where

$$\rho_N(X_i,t)=\frac{1}{h_N^2}\iint_{|y|<B_N}\mathcal{K}^*(\frac{t_n-t}{h_N})y\,dB^0(\mathcal{T}(t,y))\ , \quad (28)$$

with $\mathcal{T}:\mathbb{R}^2\to[0,1]^2$ a distribution transformation defined in [27], and B^0 is a sample path of 2D Brownian bridge; and

$$e_N(X_i,t)=-\frac{1}{\sqrt{N}h_N^2}\int\Big(\int_{|y|<B_N}y\,d_y\big[Z_N(z,y)-B^0(\mathcal{T}(z,y))\big]\Big)\,d_z\mathcal{K}^*(\frac{z-t}{h_N})\ . \quad (29)$$

with $Z_N(z,y)=\sqrt{N}\big(f_N(z,y)-f(z,y)\big)$ as a 2D-empirical process, where $f(\cdot,\cdot)$ is the distribution function of (t_n,U_i^n). It is proved in [5] that for any γ:

$$\mathbb{P}\Big[\sup_{z,y}|Z_N(t,y)-B^0(\mathcal{T}(z,y))|>\frac{(C\log N+\gamma)\log N}{\sqrt{N}}\Big]<Qe^{-L\gamma}\ ,$$

where C, Q, and L are absolute positive constants. Therefore, for $\varepsilon_{2,N}=\frac{2B_N(C\log N+\gamma)\log N}{h_N^2N}$, by change of variable in (29), we obatin:

$$\mathbb{P}\Big[\sup_t|e_N(X_i,t)|>\varepsilon_{2,N}\Big]\le\mathbb{P}\Big[\sup_t\frac{2B_N}{h_N^2\sqrt{N}}\sup_{z,y}|Z_N(z,y)-B^0(\mathcal{T}(z,y))|>\varepsilon_{2,N}\Big]<Qe^{-L\gamma}\ .$$

As for (28), similarly to (7) of [6], we have:

$$\begin{aligned}\frac{h_N^{3/2}\sup_t|\rho_N(X_i,t)|}{\sqrt{\ln\frac{1}{h_N}}}&\le 16(\ln V)^{1/2}S^{1/2}\Big(\ln\frac{1}{h_N}\Big)^{-1/2}\int|\zeta|^{1/2}\,|d\mathcal{K}^*(\zeta)|\\&\quad+16\sqrt{2}h_N^{-1/2}\Big(\ln\frac{1}{h_N}\Big)^{-1/2}\int q(Sh_N|\zeta|)\,|d\mathcal{K}^*(\zeta)|\ ,\end{aligned}$$

where V is a random variable satisfying $EV\le 4\sqrt{2}\eta^4$ for $\eta^2=\max_{n=0,\cdots,N-1}E(U_i^n)^2$, $q(r)=\int_0^r\frac{1}{2}(\frac{1}{y}\log\frac{1}{y})^{1/2}\,dy$, $S=\sup_t\int y^2f(y,t)\,dy$. From $\overline{C}=16S^{1/2}\int|\zeta|^{1/2}|d\mathcal{K}^*(\zeta)|$, for any $\varepsilon_{3,N}>2\overline{C}\sqrt{\frac{2\ln 1/h_N}{h_N^3N}}$, then we can bound the probability:

$$\mathbb{P}\Big[\sup_t|N^{-1/2}\rho_N(X_i,t)|>\varepsilon_{3,N}\Big]\le 4\sqrt{2}\eta^4\exp\frac{-N\varepsilon_{3,N}^2h_N^3}{4\overline{C}^2\ln 1/h_N}\ ,$$

when N is large. Combining the results (27), (28), and (29), for any

$$\varepsilon_N^* > \max\{2|C^*|h_N^2, \frac{12\mathcal{K}^*_{\max}B_N}{Nh_N^2}, 12AB_N^{1-s}, \frac{12B_N(C\log N+\gamma)\log N}{h_N^2 N}, 12\overline{C}\sqrt{\frac{2\ln 1/h_N}{h_N^3 N}}\} ,$$

$$\mathbb{P}\Big[\sup_{t\in[0,T]} |\Delta u_t(X_i,t)| > \varepsilon_N^*\Big] < 2N\exp(-\frac{C_N^2}{2\sigma^2}) + Qe^{-L\gamma} + 4\sqrt{2}\eta^4\exp(-72N) .$$

□

Remark A.1. In [6], there is an additional condition for B_N that, $\sum_N B_N^{-s} < \infty$. This is to guarantee that the supremum in (27) is bounded by B_N^{1-s} with probability 1. In our case, we focus on convergence in probability, hence we do not need this requirement.

Remark A.2. From [4], we obtain the asymptotic bias of this estimator:

$$E\{\widehat{u_t}(X_i,t)\} - u_t(X_i,t) = C^*(X_i)h_N^2 .$$

for some constant $C^*(X_i)$ that depends on the space point X_i. By [5, 6], under the assumptions stated in Section 3.1, we have:

$$\sup_{t\in[0,T]} |\widehat{u_t}(X_i,t) - E\widehat{u_t}(X_i,t)| = O_P\Big(\sqrt{\frac{-\log h_N}{Nh_N}}\Big) .$$

Therefore, by triangular inequality:

$$\sup_{t\in[0,T]} |\Delta u_t(X_i,t)| = O_P\Big(h_N^2 + \sqrt{\frac{-\log h_N}{Nh_N}}\Big) .$$

This expression gives the asymptotic order the error term. Our proof above provides further details of this convergence behavior.

A.2 Sufficient conditions for bounding $(\widehat{\mathbf{F}} - \mathbf{F})\boldsymbol{\beta}^*$

For the first order partial derivative estimators, we have results similarly to Proposition A.1.

Proposition A.2. *For any fixed $n = 0, \cdots, N-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist finite positive constants $A_p(t_n), \overline{C}_p(t_n), C, L, Q$ which do not depend on the spatial sample size M, such that for any $\gamma > 0$ and ε_M^* satisfying:*

$$\varepsilon^*_{M,p}(t_n) >$$

$$\max\{2|C_p^*(t_n)|w_{M,p}^2, \frac{12\mathcal{K}^*_{\max}B_M}{Mw_{M,p}^{1+p}}, 12A_p(t_n)B_M^{1-s}, \frac{12B_M(C\log M+\gamma)\log M}{w_{M,p}^{1+p}M}, 12\overline{C}_p(t_n)\sqrt{\frac{2\ln 1/w_M}{w_M^{2+p}M}}\} ,$$

where B_M is an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$, we have:

$$\mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\widehat{\partial_x^p u}(x,t_n) - \partial_x^p u(x,t_n)| > \varepsilon^*_{M,p}\Big] < 2M\exp(-\frac{C_M^2}{2\sigma^2}) + Qe^{-L\gamma} + 4\sqrt{2}\eta^4\exp(-72M) ,$$

where $C_M = B_M - ||u||_{L^\infty(\Omega)}$.

As for the product terms:

Proposition A.3. *For any fixed $n = 0, \cdots, N-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist constants L and Q which do not depend on the spatial sample size M, such that for any $\gamma > 0$ and ε^*_M satisfying:*

$$\varepsilon^*_{M,p,q} > \max\{3||\partial_x^p u(\cdot, t_n)||_\infty \varepsilon^*_{M,p},\ 3||\partial_x^q u(\cdot, t_n)||_\infty \varepsilon^*_{M,q},\ 3(\varepsilon^*_{M,p})^2,\ 3(\varepsilon^*_{M,q})^2\}$$

where B_M is an arbitrary increasing sequence $B_M \to \infty$ as $M \to \infty$; $\varepsilon^*_{M,p}$ and $\varepsilon^*_{M,q}$ are the thresholds in Proposition A.2 for the sup-norm bound of the estimator $\widehat{\partial_x^p u}$ and $\widehat{\partial_x^q u}$, respectively, $p, q = 0, 1, \cdots$, we have:

$$\begin{aligned}
&\frac{1}{4}\mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon^*_{M,p,q}\Big] \\
&< 2M\exp\Big(-\frac{C_M^2}{2\sigma^2}\Big) + Q_{p,q}(t_n)e^{-L_{p,q}(t_n)\gamma} + 4\sqrt{2}\eta^4\exp(-72M)\ ,
\end{aligned}$$

where $C_M = B_M - ||u||_{L^\infty(\Omega)}$.

Proof. Notice that for any $\varepsilon > 0$, we can bound the probability:

$$\begin{aligned}
&\mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon\Big] \\
&\le \mathbb{P}\Big[||\partial_x^p u(\cdot,t_n)||_\infty \sup_{x\in[0,X_{\max}]} |\Delta\partial_x^q u(x,t_n)| > \varepsilon/3\Big] \\
&+ \mathbb{P}\Big[||\partial_x^q u(\cdot,t_n)||_\infty \sup_{x\in[0,X_{\max}]} |\Delta\partial_x^p u(x,t_n)| > \varepsilon/3\Big] \\
&+ \mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\Delta\partial_x^p u(x,t_n)| > \sqrt{\frac{\varepsilon}{3}}\Big] + \mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\Delta\partial_x^q u(x,t_n)| > \sqrt{\frac{\varepsilon}{3}}\Big],
\end{aligned}$$

hence the results follow from Proposition A.2. □

A.3 Simplification on the Probability Bounds

We focus on the simplification of (26), and the others follow similarly. Let $a > 0$ and $b > 0$ be some positive real numbers to be specified, then the bandwidth h_N and the arbitrary increasing B_N take the forms:

$$h_N = \frac{1}{N^a},\ B_N = N^b\ . \tag{30}$$

Note that the threshold ε^*_N is determined by the maximum of 5 terms, which can be written as follows after plugging the forms in (30):

$$\begin{aligned}
&E_1(N) = \frac{2|C^*(X_i)|}{N^{2a}},\quad E_2(N) = \frac{12K^*_{\max}}{N^{1-b-2a}},\quad E_3(N) = \frac{12A(X_i)}{N^{b(s-1)}} \\
&E_4(N) = \frac{12(C\ln N + \gamma)\ln N}{N^{1-b-2a}},\quad E_5(N) = 12\overline{C}(X_i)\sqrt{\frac{2a\ln N}{N^{1-3a}}}\ .
\end{aligned} \tag{31}$$

In order to have $E_i(N) \to 0$ as $N \to \infty$, it is thus sufficient to have that:

$$\begin{cases} 1 - b - 2a > 0 \\ s - 1 > 0 \\ 1 - 3a > 0 \end{cases}$$

Notice that the probability bound in (26) becomes:

$$2N\exp(-\frac{(N^b - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) + Qe^{-L\gamma} + 4\sqrt{2}\eta^4\exp(-72N)\,, \tag{32}$$

and the free parameter γ is related to both probability upper-bound (32) and the threshold lower-bound (31). We take for some $c > 0$:

$$\gamma = \frac{N^c}{L}\,.$$

The optimal choice for such a c satisfies $c \geq \min\{2b, 1\}$. When N is sufficiently large, to determine ε_N^*, we only need to focus on comparing the powers of N in $E_i(N)$, $i = 1, 2, \cdots, 6$; this immediately leads to:

$$E_2(N) \lesssim E_4(N) \text{ and } E_3(N) \lesssim E_4(N)\,, \tag{33}$$

hence it's sufficient to only consider $E_1(N)$, $E_4(N)$, and $E_5(N)$. The optimal choice of a and b is determined by requiring that b is maximized when the following constraints are satisfied:

$$\begin{cases} 2a = 1 - b - 2a - c \\ 2a = \frac{1-3a}{2} \\ c \geq \min\{2b, 1\} \end{cases} \implies \begin{cases} a = \frac{1}{7} \\ b = \frac{1}{7} \\ c = \frac{2}{7} \end{cases} \tag{34}$$

To summarize the discussion above, we have

Corollary A.1. *For any fixed $i = 0, \cdots, M-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist finite constants $C(X_i)$ independent of N, such that with sufficiently large N and a bandwidth $h_N = N^{-1/7}$, for any $\varepsilon_N^*(X_i)$ satisfying*

$$\varepsilon_N^*(X_i) > \frac{C(X_i)\sqrt{\ln N}}{N^{2/7}}\,,$$

we have:

$$\mathbb{P}\Big[\sup_{t\in[0,T]} |\Delta u_t(X_i, t)| > \varepsilon_N^*(X_i)\Big] < 2N\exp(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2})\,. \tag{35}$$

Proof. With the choice of a, b, c in (34), the probability bound (32) becomes:

$$2N\exp(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) + Qe^{-N^{2/7}} + 4\sqrt{2}\eta^4\exp(-72N) \sim 2N\exp(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2})\,.$$

Taking $C(X_i) = \max\{2|C^*(X_i)|, 12, \sqrt{\frac{288}{7}}\overline{C}(X_i)\}$ and omitting the second order term $(\ln N/N^{2/7})^2$ gives the result. ☐

Using similar strategy, we derive analogous results for the spatial derivative errors.

Corollary A.2. *For any nonnegative integer $p = 0, 1, 2, \cdots$ and any fixed $n = 0, \cdots, N-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist finite constants $C_p(t_n) > 0$ independent of M, such that with sufficiently large M and a bandwidth $w_{M,p} = M^{-1/(6+p)}$, for any $\varepsilon^*_{M,p}$ satisfying:*

$$\varepsilon^*_{M,p}(t_n) > \frac{C_p(t_n)\sqrt{\ln M}}{M^{2/(6+p)}}\,,$$

we have:

$$\mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\widehat{\partial_x^p u}(x, t_n) - \partial_x^p u(x, t_n)| > \varepsilon^*_{M,p}\Big] < 2M\exp(-\frac{(M^{1/(6+p)} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}).$$

Corollary A.3. *For any nonnegative integers $p, q = 0, 1, 2, \cdots$ and any fixed $n = 0, \cdots, N-1$, denote $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. There exist finite constants $C_{p,q}(t_n) > 0$ independent of M, such that with sufficiently large M and a bandwidth $w_{M,p,q} = M^{-1/(6+\max(p,q))}$, for any $\varepsilon^*_{M,p,q}$ satisfying:*

$$\varepsilon^*_{M,p,q} > C_{p,q}(t_n) M^{-2/(6+\max(p,q))}$$

we have:

$$\begin{aligned}&\frac{1}{4}\mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\widehat{\partial_x^p u}(x,t_n)\widehat{\partial_x^q u}(x,t_n) - \partial_x^p u(x,t_n)\partial_x^q u(x,t_n)| > \varepsilon^*_{M,p,q}\Big] \\ &< 2M\exp(-\frac{(M^{1/(6+\max\{p,q\})} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2})\ .\end{aligned}$$

A.4 ℓ_∞ Bound for the PDE Estimation Error

Notice that in the previous results, the constants $C(X_i)$ (independent of N), $C_p(t_n)$, $C_{p,q}(t_n)$ (independent of M) for the lower bound of the error thresholds depend on the spatial or temporal grid points. To guarantee that as both $N, M \to \infty$, these constants are uniformly bounded, we prove the following lemma.

Lemma A.2. *For any integer $M \geq 1$, and any $i = 0, 1, \cdots, M-1$, $|C^*(X_i)|$ and $\overline{C}(X_i)$ in Corollary A.1 are bounded by constants that are independent of M. That is, there exist constants $C^*, \overline{C}(\sigma, ||u||_{L^\infty(\Omega)}) > 0$ such that for any $M \geq 1$*

$$\max_{i=0,\cdots,M-1} |C^*(X_i)| \leq C^*||\partial_t^3 u||_\infty, \text{ and } \max_{i=0,\cdots,M-1} \overline{C}(X_i) \leq \overline{C}(\sigma, ||u||_{L^\infty(\Omega)})\ .$$

Proof. From (3.7) in the Theorem 3.1 of [4], we see that

$$|C^*(X_i)| \leq C^*||\partial_t^3 u||_\infty < \infty$$

where C^* only depends on the choice of the kernel function and the order of the local polynomial. For a general s, we know that

$$\begin{aligned}\sup_{t\in[0,T]} \int |y|^s f(t,y|X_i)\, dy &= \sup_{t\in[0,T]} \int |y|^s \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-u(X_i,t))^2}{2\sigma^2})\, dy \\ &= \sup_{t\in[0,T]} \sigma^s 2^{s/2} \frac{\Gamma\left(\frac{1+s}{2}\right)}{\sqrt{\pi}} {}_1F_1(-\frac{s}{2}, \frac{1}{2}, -\frac{1}{2}\Big(\frac{u(X_i,t)}{\sigma}\Big)^2)\end{aligned}$$

where ${}_1F_1(a,b,z)$ is Kummer's confluent hypergeometric function of $z \in \mathbb{C}$ with parameters $a, b \in \mathbb{C}$ (See, e.g.[28]). Since ${}_1F_1(-\frac{s}{2}, \frac{1}{2}, \cdot)$ is an entire function for fixed parameters,

$$\sup_{t\in[0,T]} \int |y|^s f(t,y|X_i)\, dy \leq \sup_{t\in[0,T]} \sigma^s 2^{s/2} \frac{\Gamma\left(\frac{1+s}{2}\right)}{\sqrt{\pi}} \sup_{z\in[-\frac{\max_{x\in\Omega} u^2(x,t)}{2\sigma^2}, -\frac{\min_{x\in\Omega} u^2(x,t)}{2\sigma^2}]} {}_1F_1(-\frac{s}{2}, \frac{1}{2}, z) < \infty$$

which clearly does not depend on M. Take $s = 2$, we can obtain that $\overline{C}(X_i) \leq \overline{C}(\sigma, ||u||_{L^\infty(\Omega)})$ for some $\overline{C}(\sigma, ||u||_{L^\infty(\Omega)})$ that only depends on $||u||_{L^\infty(\Omega)}$ and σ. $\square$

Note that the same proof can derive that the constants in Proposition A.2 are also bounded by N-independent constants. This technical lemma allows us to state our main theorem:

Theorem A.1. *Denote* $\eta^2 = \max_{n=0,\cdots,N-1} E(U_i^n)^2$, $\mathcal{K}^*_{\max} = ||\mathcal{K}^*||_\infty$. *Suppose the maximal order of spatial partial derivative we consider is* $P_{\max} > 0$. *There exist some constant* $C' = C'(||u||_{W^{P_{\max},\infty}}, s, \sigma, ||\beta||_\infty, K)$ *independent of* N *and* M*, such that with sufficiently large* N *and* M*, using bandwidths* $h_N = N^{-1/7}$ *in the time direction and* $w_M = M^{-1/(6+P_{\max})}$ *in the space direction, for any* $\varepsilon_{N,M}$ *satisfying:*

$$\varepsilon_{N,M} > C' \max\{\frac{\sqrt{\ln M}}{M^{2/(6+P_{\max})}}, \frac{\sqrt{\ln N}}{N^{2/7}}\}$$

we have

$$\mathbb{P}\Big[||\boldsymbol{\tau}||_\infty > \varepsilon_{N,M}\Big] < 8MNK \exp(-\frac{(M^{1/(6+P_{\max})} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) + 2MN \exp(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) \,.$$

Here $||u||_{W^{P_{\max},\infty}(\Omega)}$ *denotes the Sobolev* $P_{\max},\infty$*-norm of* u*,* s *the data regularity parameter in Section 3.1,* σ *the noise level,* $||\boldsymbol{\beta}^*||_\infty$ *the maximal absolute coefficient, and* K *the number of feature variables.*

Proof. By triangle inequality:

$$||\boldsymbol{\tau}||_\infty \le ||\Delta \mathbf{F} \boldsymbol{\beta}^*||_\infty + ||\Delta \mathbf{u}_t||_\infty \,.$$

Since

$$||\Delta \mathbf{u}_t||_\infty \le \max_{i=0,1,\cdots,M-1} \sup_{t\in[0,T]} |\Delta u_t(X_i, t)| \,,$$

then for sufficiently large N, we have that if $\frac{\varepsilon_{N,M}}{2} > C_1' \ln N / N^{2/7}$ for some constant C_1':

$$\begin{aligned}
\mathbb{P}\Big[||\Delta \mathbf{u}_t||_\infty > \frac{\varepsilon_{N,M}}{2}\Big] &\le \mathbb{P}\Big[\max_{i=0,1,\cdots,M-1} \sup_{t\in[0,T]} |\Delta u_t(X_i,t)| > \frac{\varepsilon_{N,M}}{2}\Big] \\
&\le \sum_{i=0}^{M-1} \mathbb{P}\Big[\sup_{t\in[0,T]} |\Delta u_t(X_i,t)| > \frac{\varepsilon_{N,M}}{2}\Big] \\
&\le 2MN \exp(-\frac{(N^{1/7} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) \,.
\end{aligned}$$

On the other hand, if we denote $\Delta F_k(x,t)$ as the approximation error of the k-th feature variable at time t and space x, we have

$$||\Delta \mathbf{F} \boldsymbol{\beta}^*||_\infty \le \max_{n=0,1,\cdots,N} ||\beta||_\infty \sup_{x\in[0,X_{\max}]} \sum_{k=1}^{K} |\Delta F_k(x, t_n)| \,.$$

Hence, for sufficiently large M and if $\frac{\varepsilon_{N,M}}{2K||\boldsymbol{\beta}^*||_\infty} > C_2' \ln M / M^{2/(6+P_{max})}$ for some constant C_2':

$$\begin{aligned}
\mathbb{P}\Big[||\Delta \mathbf{F} \boldsymbol{\beta}^*||_\infty > \frac{\varepsilon_{N,M}}{2}\Big] &\le \sum_{n=0}^{N-1} \sum_{k=1}^{K} \mathbb{P}\Big[\sup_{x\in[0,X_{\max}]} |\Delta F_k(x,t_n)| > \frac{\varepsilon_{N,M}}{2K||\boldsymbol{\beta}^*||_\infty}\Big] \\
&\le 8NMK \exp(-\frac{(M^{1/(6+\max\{p,q\})} - ||u||_{L^\infty(\Omega)})^2}{2\sigma^2}) \,.
\end{aligned}$$

Taking $C' = \max\{2C_1', 2K||\boldsymbol{\beta}^*||_\infty C_2'\}$ proves the theorem. □

A.5 Final Step of the Proof

Combining Lemma A.1 with Theorem A.1, for any ε, if it satisfies:

$$\lambda\varepsilon > C'\sqrt{K}\max\{\frac{\sqrt{\ln M}}{M^{2/(6+P_{\max})}},\frac{\sqrt{\ln N}}{N^{2/7}}\}$$

then (25) implies

$$\mathbb{P}\Big[\max_{j\in\mathcal{S}^c}|\widetilde{Z}_j|>\varepsilon\Big]\leq 8MNK\exp(-\frac{(M^{1/(6+P_{\max})}-||u||_{L^\infty(\Omega)})^2}{2\sigma^2})+2MN\exp(-\frac{(N^{1/7}-||u||_{L^\infty(\Omega)})^2}{2\sigma^2})\ .$$

Therefore, if we take $\varepsilon=\mu$, then the following inequality

$$\mathbb{P}\Big[||\check{\mathbf{z}}_{\mathcal{S}^c}||_\infty\geq 1\Big]\leq\mathbb{P}\Big[||\widehat{\mathbf{F}}^T_{\mathcal{S}^c}\widehat{\mathbf{F}}_{\mathcal{S}}(\widehat{\mathbf{F}}^T_{\mathcal{S}}\widehat{\mathbf{F}}_{\mathcal{S}})^{-1}||_\infty\geq 1-\mu\Big]+\mathbb{P}\Big[\max_{j\in\mathcal{S}^c}|\widetilde{Z}_j|>\mu\Big]$$

finishes the proof of Theorem 3.1.

B Proof of Lemma 4.1

Proof. For any grid point (x_i,t_n) on $\Omega=[0,X_{\max}]\times[0,T_{\max}]$, $i=0,1,\cdots,M-1$, $n=0,1,\cdots,N-1$, if $0<x_i<X_{\max}$ and $0<t_n<T_{\max}$, it is called an inner grid point; otherwise, a boundary grid point. Now denote L as the number of rectangles centering at the inner grid points in Ω with width Δx, height Δt; and L' as that of the boundary grid points. We decompose:

$$\int_\Omega\partial_x^k u(x,t)\,dx\,dt=\underbrace{\sum_{l=1,2,\cdots,L}\int_{\Omega_l}\partial_x^k u(x,t)\,dx\,dt}_{I}+\underbrace{\sum_{l'=1,2,\cdots,L'}\int_{\Omega_{l'}}\partial_x^k u(x,t)\,dx\,dt}_{I'}\ .$$

For the first integral, by triangle inequality and Taylor's theorem:

$$\begin{aligned}&\left|I-\sum_{l=1,2,\cdots,L}\int_{\Omega_l}\partial_x^k u(x_l,t_l)\,dx\,dt\right|\\&\leq\sum_{l=1,2,\cdots,L}\int_{\Omega_l}|\partial_x^{k+1}u(\zeta_l,\eta_l)(x-\zeta_l)|\,dx\,dt+\sum_{l=1,2,\cdots,L}\int_{\Omega_l}|\partial_x^k\partial_t u(x_l,t_l)(t-\eta_l)|\,dx\,dt\\&\leq||\partial_x^{k+1}u||_{L^\infty(\Omega)}L\Delta x^2\Delta t+||\partial_x^k\partial_t u||_{L^\infty(\Omega)}L\Delta t^2\Delta x\ .\end{aligned}$$

Here (ζ_l,η_l) denote some point in the domain $(x_l-\Delta x/2,x_l+\Delta x/2)\times(t_l-\Delta t/2,t_l+\Delta t/2)$, $l=1,2,\cdots,L$. Similarly, for the second integral, we obtain:

$$\left|I'-\sum_{l'=1,2,\cdots,L'}\int_{\Omega_{l'}}\partial_x^k u(x_{l'},t_{l'})\,dx\,dt\right|\leq\frac{1}{2}(||\partial_x^{k+1}u||_{L^\infty(\Omega)}L'\Delta x^2\Delta t+||\partial_x^k\partial_t u||_{L^\infty(\Omega)}L'\Delta t^2\Delta x)\ .$$

Note that $L=(M-1)(N-1)$ and $L'=2M+2N-4$, hence

$$\begin{aligned}&\left|\int_\Omega\partial_x^k u(x,t)\,dx\,dt-\mathbf{1}^T F_k\Delta x\Delta t\right|\\&\leq||\partial_x^{k+1}u||_{L^\infty(\Omega)}|\Omega|\Delta x+||\partial_x^k\partial_t u||_{L^\infty(\Omega)}|\Omega|\Delta t\\&+\frac{1}{2}(||\partial_x^{k+1}u||_{L^\infty(\Omega)}(2X_{\max}\Delta x\Delta t+2T_{\max}\Delta x^2)+||\partial_x^k\partial_t u||_{L^\infty(\Omega)}(2X_{\max}\Delta t^2+2T_{\max}\Delta x\Delta t))\\&=o(\Delta x)+o(\Delta t)\quad\text{as }N,M\to\infty\ .\end{aligned}$$

The statement for the product feature is similarly proved once we note the following inequality:

$$
\begin{aligned}
&\left| \sum_{l=1,2,\cdots,L} \int_{\Omega_l} \partial_x^k u(x,t) \partial_x^j u(x,t)\, dx\, dt - \sum_{l=1,2,\cdots,L} \int_{\Omega_l} \partial_x^k u(x_l,t_l) \partial_x^j u(x_l,t_l)\, dx\, dt \right| \\
&\leq \sum_{l=1,2,\cdots,L} \int_{\Omega_l} |\partial_x^k u(x,t) - \partial_x^k u(x_l,t_l)|\, |\partial_x^j u(x,t)|\, dx\, dt \\
&+ \sum_{l=1,2,\cdots,L} \int_{\Omega_l} |\partial_x^k u(x_l,t_l)|\, |\partial_x^j u(x,t) - \partial_x^j u(x_l,t_l)|\, dx\, dt\ .
\end{aligned}
$$

□

References

[1] Hua Liang and Hulin Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

[2] Jianwei Chen and Hulin Wu. Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to hiv-1 dynamics. *Journal of the American Statistical Association*, 103(481):369–384, 2008.

[3] Jianwei Chen and Hulin Wu. Estimation of time-varying parameters in deterministic dynamic models. *Statistica Sinica*, 18(3):987–1006, 2008.

[4] Jianqing Fan, Theo Gasser, Irène Gijbels, Michael Brockmann, and Joachim Engel. Local polynomial regression: optimal kernels and asymptotic minimax efficiency. *Annals of the Institute of Statistical Mathematics*, 49(1):79–99, 1997.

[5] G Tusnády. A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica*, 8(1):53–55, 1977.

[6] Yue-pok Mack and Bernard W Silverman. Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 61(3):405–415, 1982.

[7] Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2197):20160446, 2017.

[8] Sung Ha Kang, Wenjing Liao, and Yingjie Liu. Ident: Identifying differential equations with numerical time evolution. *arXiv preprint arXiv:1904.03538*, 2019.

[9] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE transactions on information theory*, 55(5):2183–2202, 2009.

[10] Pradeep Ravikumar, Martin J Wainwright, John D Lafferty, et al. High-dimensional ising model selection using ?1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[11] Pradeep Ravikumar, John Lafferty, Han Liu, and Larry Wasserman. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(5):1009–1030, 2009.

[12] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, Bin Yu, et al. High-dimensional covariance estimation by minimizing ?1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[13] Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Union support recovery in high-dimensional multivariate regression. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 21–26. IEEE, 2008.

[14] Weiguang Wang, Yingbin Liang, and Eric Xing. Block regularized Lasso for multivariate multi-response linear regression. In *Artificial Intelligence and Statistics*, pages 608–617, 2013.

[15] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in neural information processing systems*, pages 964–972, 2010.

[16] VM Tikhomirov. A study of the diffusion equation with increase in the amount of substance, and its application to a biological problem. In *Selected works of AN Kolmogorov*, pages 242–270. Springer, 1991.

[17] Alan C Newell. *Solitons in mathematics and physics*, volume 48. Siam, 1985.

[18] Christopher G Fox, Haruyoshi Matsumoto, and Tai-Kwan Andy Lau. Monitoring pacific ocean seismicity from an autonomous hydrophone array. *Journal of Geophysical Research: Solid Earth*, 106(B3):4183–4206, 2001.

[19] Panuccio Michele, Stanzione Viviana, Catoni Carlo, Santini Mauro, et al. Radar tracking reveals influence of crosswinds and topography on migratory behavior of european honey buzzards. *Journal of ethology*, 34(1):73–77, 2016.

[20] Jean-Jacques Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.

[21] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.

[22] Sara A Van De Geer, Peter Bühlmann, et al. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[23] Ming-Yen Cheng, Jianqing Fan, James S Marron, et al. On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708, 1997.

[24] Trevor Hastie, Clive Loader, et al. Local regression: Automatic kernel carpentry. *Statistical Science*, 8(2):120–129, 1993.

[25] Richard Haberman. *Elementary applied partial differential equations*, volume 987. Prentice Hall Englewood Cliffs, NJ, 1983.

[26] Andreas M Tillmann and Marc E Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2013.

[27] Murray Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.

[28] Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *arXiv preprint arXiv:1209.4340*, 2012.