Asymptotic Theory of ℓ_1-Regularized PDE Identification from a Single Noisy Trajectory *

Yuchen He [†‡], Namjoon Suh [†§], Xiaoming Huo [§], Sung Ha Kang[¶], and Yajun Mei[§]

Abstract. We provide a formal theoretical analysis on the PDE identification via ℓ_1-regularized Pseudo Least Square method from the statistical point of view. In this article, we assume that the differential equation governing the dynamic system can be represented as a linear combination of various linear and nonlinear differential terms. Under noisy observations, we employ the Local-Polynomial fitting for estimating state variables and apply the ℓ_1 penalty for model selection. Our theory proves that the classical mutual incoherence condition on the feature matrix $\mathbf{F}$ and $\boldsymbol{\beta}^*_{\min}$-condition for the ground-truth signal $\boldsymbol{\beta}^*$ are sufficient for the signed-support recovery of ℓ_1-PsLS method. We run numerical experiments on two popular PDE models, viscous Burgers and KdV equations, and the results from the experiments corroborate our theoretical predictions.

Key words. Parital Differential Equation (PDE), Lasso, Pseudo Least Square, Signed-Support Recovery, Primal-Dual Witness construction, Local-Polynomial Regression.

AMS subject classifications. 62J07, 93B30, 35G35

1. Introduction.

Differential equations are widely used to describe many interesting phenomena arising in scientific fields, including physics [1], social sciences [2], biomedical sciences [3], and economics [4], just to name a few. The forward problem of solving equations or simulating state variables for differential models has been extensively studied either theoretically or numerically in literature. We consider an inverse problem of learning a Partial Differential Equations (PDE) model.

More specifically, we assume that the governing PDE is a multi-variate polynomial of a subset of a prescribed dictionary containing different differential terms. Let $u(x,t) : \mathbb{R} \times [0,+\infty) \to \mathbb{R}$ be a real-valued function, where x be the spatial and t be the temporal variables. Suppose that within a bounded region of $\mathbb{R} \times [0,+\infty)$, $u(x,t)$ satisfies an evolutionary PDE:

$$\partial_t u = \mathcal{F}(u, \partial_x u, \partial_x^2 u, \dots,), \quad \forall (x,t) \in \Omega \subseteq \mathbb{R} \times [0,+\infty). \tag{1.1}$$

Here, $\partial_t u$ (or u_t) denotes the partial derivative of u with respect to temporal variable, t; for $p = 0, 1, 2, \dots, \partial_x^p u$ denotes the p-th order partial derivative of u with respect to spatial variable, x; $\mathcal{F}$ is an unknown polynomial mapping, and Ω is a bounded open subset of space-time

[†]Yuchen He and Namjoon Suh contributed equally to this article, listed in alphabetical order.

[*]Submitted to the editors DATE.

Funding: This project is partially supported by the Transdisciplinary Research Institute for Advancing Data Science (TRIAD), http://triad.gatech.edu, which is a part of the TRIPODS program at NSF and locates at Georgia Tech, enabled by the NSF grant CCF-1740776. X. Huo's research was supported in part by NSF grant DMS-2015363. S. H. Kang's research was supported in part by Simons Foundation Collaboration Grants 584960. Y. Mei's research was supported in part by NSF-DMS grant 2015405.

[‡]Institute of Natural Sciences, Shanghai Jiao Tong University, Shanghai, China

[§]School of Industrial and Systems Engineering, Georgia Institute of Technology, 755 Ferst Dr, Atlanta, GA, USA.

[¶]School of Mathematics, Georgia Institute of Technology, 686 Cherry St NW, Atlanta, GA, USA.

domain. 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 [5], Kolmogorov-Petrovsky-Piskunov equation describing phase transitions [6], and Korteweg-de-Vries equation simulating the shallow water dynamics [7].

In our work, $\mathcal{F}$ is assumed to be a linear map, parametrized by a sparse vector $\boldsymbol{\beta}^* \in \mathbb{R}^K$: that is, u_t is represented as a linear combination of the arguments of $\mathcal{F}$, and only a few from a large set of potential functions are relevant with u_t. Our goal is to estimate the correct non-zero indices of $\boldsymbol{\beta}^*$, given a single noisy trajectory of the function $u(x,t)$. Readers can refer to Subsection 3.1 for more detailed descriptions on the structural assumptions on $\mathcal{F}$, $\boldsymbol{\beta}^*$, and noisy trajectory. This problem setting naturally leads us to develop a two-stage method for the PDE identification based on Local-Polynomial smoothing and the ℓ_1-regularized Pseudo Least Square (ℓ_1-PsLS) method. In the first stage, from a given noisy observation, we propose to estimate the underlying bi-variate function $u(x,t)$ and its partial derivatives with respect to its spatial and temporal dimensions via the Local-Polynomial fitting [8, 9]. In the second stage, with the constructed functions through Local-Polynomial regression, we propose to identify the correct differential terms and estimate model parameters via an ℓ_1-regularized Pseudo Least-Square method.

We note that the two-stage method with Local-Polynomial regression has been applied in the Ordinary Differential Equations (ODE) setting. Specifically, the paper [10] established the consistency and asymptotic normality of the pseudo least square estimator in the ODE setting, where they used Local-Polynomial regression to estimate the state variables from the noisy data. Similarly, [11, 12] studied the parameter estimation of ODE models with varying coefficients. However, these literature focused on estimating model parameters, rather than on selecting correct differential models. In the context of PDE, [13] studied PDE identification problems, using two-stage method. Authors of the paper modeled unknown PDEs using multivariate polynomials of sufficiently high order, and the best fit was chosen by minimizing the least squares error of the polynomial approximation. Nonetheless, ℓ_1 penalization for model selection was not used, and theoretical justification for their method remains underdeveloped.

From the theoretical point of view, our paper is the first work to propose the method, ℓ_1-PsLS, with a provable guarantee in the PDE recovery problem. Our main theoretical contribution is to establish sufficient conditions for ***signed-support recovery*** of the proposed ℓ_1-PsLS in PDE identification problems. It is worth noting that the signed-support recovery is a slightly stronger criterion than the support recovery, where its primary goal is not limited to finding the non-zero indices of $\boldsymbol{\beta}^*$, but also aims at recovering the correct signs of the selected coefficients. Ensuring the correct signed-support recovery of governing dynamical system has an important practical implication since many PDEs are sensitive to the signs of coefficients. For example, changing the sign of the advection term in the transport equation reverses the moving direction, and inverting the sign of the Laplacian term of heat equation leads to instability of the system of interest.

Our theorem states that following two main conditions are sufficient for the signed-support recovery of ℓ_1-PsLS: (i) ***mutual incoherence condition*** among the arguments of the map $\mathcal{F}$, and (ii) $\boldsymbol{\beta}^*_{min}$***-condition*** on $\boldsymbol{\beta}^*$. The first condition states that a large number of irrelevant predictors cannot exhibit an overly strong influence on the subset of relevant predictors. The second condition says that the minimum absolute value of non-zero entries of $\boldsymbol{\beta}^*$ should be

greater than a certain threshold. These conditions appear in the statistical literature on the signed-support/support recovery of Lasso [14, 15, 16, 17] in linear regression problems, and our work rigorously shows that these are also essential for the signed-support recovery of PDE identification problems.

We employ Primal-Dual Witness (PDW) construction [15] as the main proof technique for the theorem. PDW construction is a popular mathematical technique for certifying variable-selection consistency of ℓ_1-penalized M-estimation problems including Lasso. See [18, 19, 20, 21, 22, 23]. For reader's convenience, we provide a brief introduction of the technique in the supplementary material SM1. However, we want to emphasize that our Theorem is not a direct result of the trivial application of the PDW construction. Our problem settings are different from those of the work [15] in two aspects, which add some delicacies to our proof:

- As will be detailed in Subsection 3.3, the distribution of residual vector $\boldsymbol{\tau}$ is unknown, and neither mean 0 nor independent in our setting. On the contrary, in the work of [15], each entry of the residual vector is assumed to follow centered Gaussian with $\sigma^2 > 0$ variance and independent with the others.

- In the ℓ_1-PsLS method, the feature matrix obtained via Local-Polynomial fitting from noisy data is always random and has dependent rows uniquely determined through the underlying PDE. On the other hand, [15] divided their analysis into two cases, where the feature matrix X is either deterministic or random. When X is random, it is assumed to be a Gaussian ensemble with independent rows, whose covariance matrix that satisfies mutual incoherence condition.

Organization. The remainder of the paper is organized as follows. Some related literature with our work are reviewed in Section 2. In Section 3, we formally define our problem by imposing some specific structural assumptions on $\mathcal{F}$ and propose a ℓ_1-PsLS method for PDE identification. In Section 4, the main theorem of our work is given on the signed-support recovery of ℓ_1-PsLS with the mutual incoherence assumption on the feature matrix $\mathbf{F}$, and we provide a high-level outline of the proof. Section 5 is devoted to provide a similar result with that of the one in the main theorem in Section 4 under milder assumption: that is, mutual incoherence assumption is imposed on the estimated feature matrix $\widehat{\mathbf{F}}$; an overview of proof is furnished. Related technical difficulties for the proof and main technical contribution of the paper are also given. Section 6 provides two Lemmas for completing the proof of the main Theorem by linking the mutual incoherence assumption with ground-truth $\mathbf{F}$ to its sampled version. In Section 7, we show various numerical examples to validate and demonstrate different aspects of our method. We conclude this paper in Section 8 with some discussion.

Notation. For sufficiently large n, we write $f(n) = \mathcal{O}(g(n))$, if there exists a constant $K > 0$ such that $f(n) \leq Kg(n)$, and $f(n) = \Omega(g(n))$ if $f(n) \geq K'g(n)$ for some constant $K' > 0$. The notation $f(n) = \Theta(g(n))$ means that $f(n) = \mathcal{O}(g(n))$ and $f(n) = \Omega(g(n))$. We adopt bold lower-case letters for vectors and bold upper-case letters for matrices. For a vector $\mathbf{v} \in \mathbb{R}^n$, $\|\mathbf{v}\|_1 := \sum_{i=1}^n |\mathbf{v}_i|$, $\|\mathbf{v}\|_2 := \sqrt{\sum_{i=1}^n \mathbf{v}_i^2}$, and $\|\mathbf{v}\|_\infty := \max_{1\leq i\leq n} |\mathbf{v}_i|$. For a matrix $\mathbf{A} \in$

$\mathbb{R}^{n\times m}$, $\mathbf{A}^T$ denotes its transpose, $|||\mathbf{A}|||_2 := \max_{\forall \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$, $|||\mathbf{A}|||_\infty := \max_{1\le i\le n} \sum_{j=1}^m |\mathbf{A}_{i,j}|$, $|||\mathbf{A}|||_{\infty,\infty} := \max_{1\le i,j\le n} |\mathbf{A}_{i,j}|$, and $|||\mathbf{A}|||_F := \sqrt{\sum_{i=1}^n \sum_{j=1}^m \mathbf{A}_{i,j}^2}$.

2. Related Works. Our work is relevant to various topics in applied mathematics and statistics. Among them, we provide two most closely related topics: (i) Regression-based framework for PDE identification, and (ii) Some theoretical results of support-recovery of Lasso [14] in linear regression setting. In this Section, we denote K as the problem dimension, s as the number of non-zero entries of model parameter, and n as the number of observations.

Regression-based Methods. Recently, various regression-based frameworks have been developed and applied for model selection and parameter estimation of dynamic data. A sparsity-promoting method was proposed in [24] for extracting the governing dynamical system, by comparing the computed velocity to a large set of potential trial functions. Under the over-determined systems of linear equations (i.e., $n \gg k$), the authors developed a sequential-thresholded least-square method to select the correct nonlinear functions. In the follow-up study, [25] devised a weighted-ℓ_1-regularized least squares solver for improving the accuracy and robustness of the approach introduced by [24] in the presence of state-measurement noise. Several papers [26, 27, 28] also suggested sparse regression frameworks for PDE identification problems over spatial-temporal data. Specifically, [27] studied the model selection problem via Lasso under the PDE context. The author empirically showed that the method works well in various important equations such as Burgers' equation, Navier-Stokes equation, Swift-Hohenberg equation. Recently, [26] 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. Unlike the previously mentioned literature, which was mostly empirical, [29] provided a provable guarantee on the usage of ℓ_1-norm for PDE identification problems, based upon the theoretical results from compressive sensing. Interestingly, this work imposed the *incoherence property* on the feature matrix and employed the Legendre-transform on the columns of the matrix to ensure that the property holds for every PDE recovery problem of interest. Our work imposes mutual incoherence assumption on the feature matrix, which is an analogous notion of the incoherence property. However, the important difference between our paper and [29] is that our work only allows a single trajectory, whereas [29]'s theorem requires $\Omega(s \log K)$ bursts of initialization for the exact recovery of the underlying PDE.

Support Recovery in Statistics. Support recovery or variable selection problems of Lasso have a long history in the statistical literature. In the noiseless setting, many researchers [30, 31, 32, 33, 34, 35] established sufficient conditions for either the deterministic or random predictors for the support recovery problems of linear systems via the ℓ_1-norm.

Since our work falls into the category of noisy setting, we focus more on reviewing the body of work in the noisy setting. In [36], authors studied the asymptotic behavior of the Lasso-type estimator with fixed dimension K under the general centered i.i.d. noises with variance $\sigma^2 > 0$. Both [37] and [32] independently developed sufficient conditions for the support of Lasso estimator to be contained within true support of the sparse model. Under

a more general setting, when the exterior noise is i.i.d. with finite moments, [38] showed that the Irrepresentable Condition [39] is almost necessary and sufficient for Lasso's signed-support recovery for fixed K and s. Furthermore, under the Gaussian noise assumption, they showed that Lasso can still achieve signed-support recovery when K is allowed to grow exponentially faster than n. In a non-asymptotic setting, [15] established the sharp relationship of n, K, and s, required for the exact sign consistency of Lasso, where K and s are allowed to grow as n increases under mutual incoherence condition. Using a similar technique in [15], the paper [16] studied Lasso under Poisson-like model with heteroscedastic noise and show that irrepresentable condition can serve as a necessary and sufficient condition for signed-support recovery in their setting. In the context of graphical model, [40, 41] analyzed the model selection consistency of Gaussian graphical models, and [18] showed the signed-support recovery of Ising models. See [42] for a more comprehensive overview on this topic.

Remark 2.1. Our work is of asymptotic nature with fixed K and s, while the number of grid points of the observed trajectory tends to infinity in both space and time.

3. PDE Identification via ℓ_1-PsLS.

In Subsection 3.1, we provide concrete problem settings on the governing PDE of (1.1) and the observed trajectory. Then, specific settings of the Local-Polynomial regression for the estimations of state variables in our paper are provided in Subsection 3.2. Lastly, we propose a two-stage ℓ_1-regularized Pseudo Least Square method for PDE identification in Subsection 3.3.

3.1. Problem Setting and Notations.

Based on the general form (1.1), we take $(x,t) \in [0, X_{\max}) \times [0, T_{\max})$ for some finite constants $0 < X_{\max}, T_{\max} < \infty$. It is assumed that the underlying mapping $\mathcal{F}$ is a degree 2 polynomial in terms of u and its partial derivatives $\partial_x^p u$ for $1 \le p \le P_{\max}$, [1] parametrized by a coefficient vector $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, \dots, \beta_{p,q}^*, \dots)$ with real entries; that is,

$$u_t(x,t) = \beta_0^* + \beta_1^* u + \beta_2^* u_x + \beta_3^* u_{xx} + \cdots + \beta_{p,q}^* \partial_x^p u \partial_x^q u + \dots . \tag{3.1}$$

We call the monomials in the right-hand side of (3.1) as *feature variables*. 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 (3.1). Hence, We assume that $\boldsymbol{\beta}^* \in \mathbb{R}^K$, with $K = 1 + 2(P_{\max}+1) + \binom{P_{\max}+1}{2}$; consequently, constant and any term of the form $\partial_x^p u$ or $\partial_x^p u \partial_x^q u$, for $0 \le p, q \le P_{\max}$, are contained in (3.1). Notice that many entries of $\boldsymbol{\beta}^*$ can be zero. We denote $\mathcal{S}(\boldsymbol{\beta}^*) := \{0 \le j \le K \mid \beta_j^* \neq 0\}$, or simply $\mathcal{S}$, as the support of the coefficient vector $\boldsymbol{\beta}^*$, i.e., the set of indices of the non-zero entries. Additionally, we denote s as the cardinality of the set $\mathcal{S}$, i.e., $s := |\mathcal{S}(\boldsymbol{\beta}^*)|$.

The given data $\mathcal{D} = \{(X_i, t_n, U_i^n) \mid \ i = 0, \dots, M-1; n = 0, \dots, N-1\} \subseteq \Omega \times \mathbb{R}$ consists of $M \times N$ data, where $M, N \in \mathbb{R}$, $M, N \ge 1$. Each $(X_i, t_n) \in \Omega$ represents a space-time point, and U_i^n is a representation of $u(X_i, t_n)$ contaminated by additive Gaussian noise:

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

[1] It should be noted that our setting can be generalized to higher-degrees of polynomials and functions with multiple spatial dimensions.

whose second moment is uniformly bounded as follows: $\sup_{N,M\in\mathbb{R}} \max_{n,i} E\,|U_i^n|^2 := \eta^2 < \infty$. Here $\mathcal{N}(0,\sigma^2)$ denotes the centered normal distribution with variance $\sigma^2 > 0$.

3.2. Local-Polynomial Regression Estimators for Derivatives. Given data $\{(X_i, t_n, U_i^n)\}$ with $i = 0, 1, \ldots, M-1$ and $n = 0, 1, \ldots, N-1$, we employ a local quadratic regression to estimate $u_t(X_i, \cdot)$ for each fixed space point X_i and use a Local-Polynomial with degree $p+1$ to estimate $\partial_x^p u(\cdot, t_n)$ at each temporal point t_n, for each degree $p = 0, 1, \ldots, P_{\max}$. More specifically, we solve the following optimization problems:

$$\left\{\widehat{b}_j(X_i,t)\right\}_{j=0,1,2} = \underset{b_j(t)\in\mathbb{R},0\le j\le 2}{\arg\min} \sum_{n=0}^{N-1} \left(U_i^n - \sum_{j=0}^{2} b_j(t)(t_n - t)^j\right)^2 \mathcal{K}_{h_N}\left(t_n - t\right), \tag{3.2}$$

for $i = 0, 1, \ldots, M-1$;

$$\left\{\widehat{c}_j^p(x,t_n)\right\}_{j=0,1,\ldots,p+1} = \underset{c_j(t)\in\mathbb{R},0\le j\le p+1}{\arg\min} \sum_{i=0}^{M-1} \left(U_i^n - \sum_{j=0}^{p+1} c_j^p(t)(X_i - x)^j\right)^2 \mathcal{K}_{w_M}\left(X_i - x\right), \tag{3.3}$$

for $n = 0, 1, \ldots, N-1$ and $p = 0, 1, \ldots, P_{\max}$.

and set $\widehat{u}_t(X_i,t) = \widehat{b}_1(X_i,t)$ and $\widehat{\partial_x^p u}(x,t_n) = p!\widehat{c}_p^p(x,t_n)$. Here h_N and $w_{p,M}$ denote the window width parameters, and $\mathcal{K}_w(z) := \mathcal{K}(z/w)/w$ for some kernel function $\mathcal{K}$ with window width $w > 0$. Specific choices of the order of polynomial fit for the functions $\widehat{u}_t$ and $\widehat{\partial_x^p u}$ are to strike the balance between modeling bias and variance. See Subsections 3.1 and 3.3 of Fan and Gijbels [8] for more rigorous treatments on this topic. Also the kernel $\mathcal{K}$ is assumed to be uniformly continuous and absolutely integrable with respect to Lebesgue measure on the real-line; $\mathcal{K}(z) \to 0$ as $|z| \to +\infty$; and $\int |z \ln |z||^{1/2} |dK(z)| < +\infty$.

Optimization problems (3.2) and (3.3) have closed-form solutions in the form of weighted least square estimator. See supplementary material SM2. However, for theoretical investigation, we employ the notion of *equivalent kernel* [8, 9] to write the solutions as follows: for any fixed spatial point X_i, $i = 0, 1, \ldots, M-1$, $\widehat{u}_t(X_i,t)$ can be written as:

$$\widehat{u}_t(X_i,t) = \frac{1}{Nh_N^2} \sum_{n=0}^{N-1} \mathcal{K}_2^*\left(\frac{t_n - t}{h_N}\right) U_i^n \{1 + o_{\mathbb{P}}(1)\}. \tag{3.4}$$

Similarly, for any fixed temporal point t_n, $n = 0, 1, \ldots, N-1$, the estimation for the p-th order partial derivative takes the form:

$$\widehat{\partial_x^p u}(x,t_n) = \frac{p!}{Mw_M^{p+1}} \sum_{i=1}^{M} \mathcal{K}_p^*\left(\frac{X_i - x}{w_M}\right) U_i^n \{1 + o_{\mathbb{P}}(1)\}. \tag{3.5}$$

Here, $\mathcal{K}_j^*(z) = e_j^\top S^{-1}(1, z, \ldots, z^p)^\top K(z)$ is called an equivalent kernel, where e_j denotes a unit vector with 1 on the j^{th} position; $S = (\int z^{l+s}\mathcal{K}(z)dz)_{0\le l,s\le p}$ is the moment matrix associated with kernel $\mathcal{K}$; and $o_{\mathbb{P}}(1)$ denotes a random quantity tending to zero as either N or M tends to infinity. From here, we will omit the dependency on j for the simplicity of notation when using the equivalent kernel.

Remark 3.1. The most important reason for using the Local-Polynomial fitting for the estimation of state variables and their derivatives is its rich literature in asymptotic properties and uniform convergence of the estimator [8, 43, 44, 9]. Specifically, these results allow us to explore the behavior of the tail-probability of the measurement error $\boldsymbol{\tau}$, which is essential for the analysis of the ℓ_1-PsLS estimator. See Subsection 5.2 for more information.

3.3. ℓ_1-regularized Pseudo Least Square Model. First, we introduce matrix-vector notations for compact expressions of the problem. We let $\mathbf{u}_t \in \mathbb{R}^{NM}$ denote the vectorization of $\{u_t(X_i, t_n)\}_{i=0,\dots,M-1}^{n=0,\dots,N-1}$ 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:

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

With these notations, (3.1) can be written as $\mathbf{u}_t = \mathbf{F}\boldsymbol{\beta}^*$. Note that before estimating the correct signed-support of $\boldsymbol{\beta}^*$, $\mathbf{u}_t$ and $\mathbf{F}$ need to be estimated. Conventional regression techniques such as Local-Polynomial regression, smoothing spline, and regression spline, among others, can be used to estimate $\mathbf{u}_t$ and columns of $\mathbf{F}$. In this paper, we employ the Local-Polynomial approach. We denote $\widehat{\mathbf{u}}_t \in \mathbb{R}^{NM}$ 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., $\widehat{(u_t)_i^n}$, $\widehat{(\partial_x^p u)_i^n}$, and $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}$.)

Let $\Delta\mathbf{u}_t = \widehat{\mathbf{u}}_t - \mathbf{u}_t$, $\Delta\mathbf{F} = \widehat{\mathbf{F}} - \mathbf{F}$ denote the difference between the obtained estimators $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ via Local-Polynomial regression and their ground-truth counterparts. With these notations, we formally obtain a regression model

$$\widehat{\mathbf{u}}_t = \widehat{\mathbf{F}}\boldsymbol{\beta}^* + \boldsymbol{\tau}\,, \quad \text{where } \boldsymbol{\tau} = \Delta\mathbf{F}\boldsymbol{\beta}^* - \Delta\mathbf{u}_t\,. \tag{3.6}$$

The natural extension for inducing sparsity of the parameter of interest is to add positively weighted ℓ_1-penalty term $\|\boldsymbol{\beta}\|_1$ to the squared loss $\|\widehat{u}_t - \widehat{F}\boldsymbol{\beta}\|_2^2$, leading to an estimator:

$$\widehat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \left\{ \frac{1}{2NM} \left\| \widehat{\mathbf{u}}_t - \widehat{\mathbf{F}}\boldsymbol{\beta} \right\|_2^2 + \lambda_N \|\boldsymbol{\beta}\|_1 \right\}, \tag{3.7}$$

where $\lambda_N > 0$ is a regularization hyper-parameter. Note that we normalize the columns of $\widehat{\mathbf{F}}$ such that $\frac{1}{\sqrt{NM}} \max_{j=1,\dots,K} \|\widehat{\mathbf{F}}_j\|_2 \le 1$ while solving (3.7).

Observe that (3.7) is formally identical to LASSO [14] for high-dimensional sparsity recovery. Meanwhile, we should also emphasize that $\widehat{\boldsymbol{\beta}}^{\lambda}$ is not a true ℓ_1-least square estimator, but a minimizer of the ℓ_1-least square fit with the estimated $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$, instead of the ground-truth $\mathbf{u}_t$ and $\mathbf{F}$. Hence, we use the word ***"pseudo"*** as in [10] to emphasize the approximations of the

solutions and derivatives of measurements, and call our method ℓ_1**-Pseudo Least Square** method. Additionally, the residual vector $\boldsymbol{\tau}$ violates conventional assumptions on residuals in linear regression, where they are assumed to be centered and independent among entries. See [38, 15, 36]. If $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ are unbiased estimators of $\mathbf{u}_t$ and $\mathbf{F}$, $\boldsymbol{\tau}$ is a residual vector with mean zero, but its entries are not independent. However, if $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ are biased estimators such as Local-Polynomial estimators in our case, $\boldsymbol{\tau}$ is not a mean zero random vector. Moreover, the unknown signal $\boldsymbol{\beta}^*$ makes the distribution of $\boldsymbol{\tau}$ completely inaccessible. These complexities make the study of the proposed estimator $\widehat{\boldsymbol{\beta}}^{\lambda}$ challenging.

4. Recovery Theory for ℓ_1-PsLS based PDE Identification. In subsection 4.1, we formally describe a signed-support recovery problem. In subsection 4.2, two regularity assumptions on feature matrix $\mathbf{F}$ are given for the proof of the main theorem. Then, the main theorem of this work is presented with some important remarks in subsection 4.3. Lastly, we provide a proof sketch of the main Theorem in subsection 4.4.

4.1. Signed-Support Recovery. The main goal of this paper is to provide provable guarantees that the proposed ℓ_1-PsLS method gives asymptotically consistent estimator of $\boldsymbol{\beta}^*$ in the sense of signed-support recovery. We can formally state this problem with the adoption of $\mathbb{S}_{\pm}(\boldsymbol{\beta})$ notation, that is: for any vector $\boldsymbol{\beta} \in \mathbb{R}^K$, we define its extended sign vector, whose each entry is written as:

$$\mathbb{S}_{\pm}(\beta_i) := \begin{cases} +1 & \text{if } \beta_i > 0 \\ -1 & \text{if } \beta_i < 0 \\ 0 & \text{if } \beta_i = 0, \end{cases}$$

for $i \in \{1, \dots, K\}$. This notation encodes the *signed-support* of the vector $\boldsymbol{\beta}$. Denote $\widehat{\boldsymbol{\beta}}^{\lambda}$ as the unique solution of ℓ_1-PsLS. Under some regularity conditions on $\mathbf{F}$, we will show,

$$\mathbb{P}\big[\mathbb{S}_{\pm}(\widehat{\boldsymbol{\beta}}^{\lambda}) = \mathbb{S}_{\pm}(\boldsymbol{\beta}^*)\big] \to 1 \quad \text{as } N, M \to +\infty,$$

where N and M denote the grid size of temporal and spatial dimensions, respectively.

4.2. Assumptions. We introduce two sufficient conditions frequently assumed in ℓ_1- regularized regression models for the signed-support recovery of the true signal $\boldsymbol{\beta}^*$.

1. **Minimal eigenvalue condition.** There exists some constant $C_{\min} > 0$ such that:

$$\Lambda_{\min}\left(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right) \geq C_{min}. \tag{A1}$$

   Here $\Lambda_{\min}(\mathbf{A})$ denotes the minimal eigenvalue of a square matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$, and $\mathbf{F}_{\mathcal{S}}$ is made of columns of $\mathbf{F}$ when the column index is in the support set $\mathcal{S}$. Note that if this condition is violated, the columns of $\mathbf{F}_{\mathcal{S}}$ would be linearly dependent, and it would be impossible to estimate the true signal β^* even in the "oracle case" when the support set $\mathcal{S}$ is known a priori.
2. **Mutual incoherence condition.** For some *incoherence parameter* $\mu \in (0, 1]$:

$$\left\| \left(\mathbf{F}_{\mathcal{S}^c}^{\top}\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right)^{-1} \right\|_{\infty} \leq 1 - \mu. \tag{A2}$$

This condition states that the irrelevant predictors cannot exhibit an overly strong influence on the relevant predictors. More specifically, for each index $j \in \mathcal{S}^c$, the vector $(\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_{\mathcal{S}})^{-1}\mathbf{F}_{\mathcal{S}}^T\mathbf{F}_j$ is the regression coefficient of $\mathbf{F}_j$ on $\mathbf{F}_{\mathcal{S}}$, thus, it is a measure of how well the column $\mathbf{F}_j$ aligns with the columns of $\mathbf{F}_{\mathcal{S}}$. A large μ close to 1 indicates that the columns $\{\mathbf{F}_j, j \in \mathcal{S}^c\}$ are nearly orthogonal to the columns of $\mathbf{F}_{\mathcal{S}}$, which is desirable for support recovery.

For future reference, we define $\mathcal{Q}^* := \left(\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}$, and name it as *population incoherence matrix*. Also, define its estimated counterpart as $\widehat{\mathcal{Q}}_N := \left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$, and call it *sample incoherence matrix*. Note that the dependence of the support set $\mathcal{S}$ on quantities $\mathcal{Q}^*$ and $\widehat{\mathcal{Q}}_N$ is suppressed for notational simplicity.

4.3. Statement of Main Result.

Theorem 4.1. *Given the observed data set $\mathcal{D}$ whose spatial resolution is related to the temporal resolution via $M = \Theta(N^{\frac{2P_{\max}+5}{7}})$, we take the bandwidths of the kernels in (3.2) and (3.3) as $h_N = \Theta(N^{-\frac{1}{7}})$, $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. Under the assumptions (A1) and (A2) imposed on the ground-truth feature matrix $\mathbf{F}$, suppose that the sequence of regularization hyper-parameters $\{\lambda_N\}$ satisfies $\lambda_N = \Omega\left(\frac{\sqrt{K}\ln N}{\mu N^{2/7-c}}\right)$ for some constant $0 < c < \frac{2}{7}$ independent of N. Then, the following properties hold with probability greater than $1 - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}}\exp\left(-\frac{1}{6}N^c\right)\right) \to 1$ as $N \to \infty$:*

(i) The ℓ_1-PsLS method (3.7) has a unique minimizer $\widehat{\boldsymbol{\beta}}^\lambda \in \mathbb{R}^K$ with its support contained within the true support, that is $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$, and the estimator satisfies the ℓ_∞ bound:

$$\left\|\widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}}\right\|_\infty \leq K^{3/2}C_{\min}\left(o_N(1) + \lambda_N\right) . \tag{4.1}$$

(ii) Additionally, if the minimum value of the model parameters supported on $\mathcal{S}$ is greater than the upper-bound of (4.1), that is $\min_{1\leq i\leq s}|(\boldsymbol{\beta}^*_{\mathcal{S}})_i| > K^{3/2}C_{\min}\left(o_N(1) + \lambda_N\right)$, then $\widehat{\boldsymbol{\beta}}^\lambda$ has a correct signed-support. i.e., $\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}^\lambda) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)$.

The overall proof sketch of Theorem 4.1 is described in the Subsection 4.4, and relevant technical propositions and Lemmas are further provided in Sections 4 and 5. Here, we give some important remarks about Theorem 4.1.

1. The uniqueness claim of $\widehat{\beta}^\lambda$ in (i) seems trivial since the objective function in (3.7) is strictly convex in the regime of K being fixed and $NM \to \infty$. However, we need to ensure that the minimal eigenvalue condition hold over the estimated feature matrix $\widehat{\mathbf{F}}$, given the assumption (A1) for some $C_{\min} > 0$. We defer this statement as Lemma 6.2 in Section 6 with the detailed proof.

2. The item (i) claims that ℓ_1-PsLS does not select the arguments that are not in the support of $\boldsymbol{\beta}^*$. The item (ii) is a consequence of the sup-norm bound from (4.1): as long as $|\boldsymbol{\beta}^*_i|$ over indices $i \in \mathcal{S}$ is not small, ℓ_1-PsLS is signed-support recovery consistent.

3. The asymptotic orders of M, h_N, and w_M are specifically chosen for simplicity. Although there is certain flexibility, the spatial resolution M and the temporal resolution N (as well as h_N and w_M) need to be coordinated well to guarantee the support recovery property. This was expected in practice since we need sufficient sampling frequencies both in temporal and space to estimate the underlying dynamics. Here, the Theorem 4.1 present a rigorous justification for a combination of these resolutions which is sufficient for the support recovery.

4. The quantity c is derived from the Tusnády's strong approximation [44] where the error of an empirical distribution is compared with a Brownian bridge in tail probability. See supplementary material SM3.1. With a larger value of c, the regularization hyper-parameter λ_N needs to remain relatively large, but the convergence is faster. Whereas for a smaller value of c, we can relax the regularization in the cost of a slower probability convergence rate.

5. The threshold of λ_N in the statement of the Theorem 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, to guarantee that the support of the estimated coefficient vector is contained in the correct one, it suffices to use a large value of λ_N. Such behavior of the threshold is consistent with that described in Theorem 1 of [15].

6. The upper-bound for the ℓ_∞-norm of the coefficient error in (4.1) consists of two components. The first term $o_N(1)$ denotes a deterministic sequence converging to 0 as N increases to ∞. We want to note that this term is involved with the underlying function u as well as the choice of regression kernels and independent with the choice of feature variables selected by ℓ_1-PsLS. The second component is simple: $K^{3/2}C_{\min}\lambda_N$. When N increases, this part does not vary. This indicates that asymptotically, ℓ_1-PsLS recovers signed-support of governing PDE, as long as $\min_{1\le i\le s}|(\boldsymbol{\beta}^*_{\mathcal{S}})_i| > K^{3/2}C_{\min}\lambda_N$.

4.4. Proof Strategy of Theorem 4.1. The analysis for the proof of Theorem 4.1 is naturally divided into two steps as follows: In the first step, we prove a result analogous to that of the Theorem 4.1 by imposing incoherence assumption on the estimated feature matrix $\widehat{\mathbf{F}}$. Specifically, since $\widehat{\mathbf{F}}$ is a random matrix, we assume that for some $\mu \in (0,1]$, the event, $\{|||\widehat{\mathcal{Q}}_N|||_\infty \le 1-\mu\}$, holds with some probability at least P_μ, for some $P_\mu \in (0,1]$. Under this assumption, we prove that the success probability of signed-support recovery of ℓ_1-PsLS converges to P_μ with an exponential decay rate. This is formally stated as Proposition 5.1 in Subsection 5.1.

In the second step, we show that the success probability P_μ goes to 1, given that the ground-truth matrix $\mathbf{F}$ satisfies assumptions (A1) and (A2). This is equivalent to proving that, given the assumptions (A1) and (A2) for $\mathbf{F}$ for some $C_{\min} > 0$ and $\mu \in (0,1]$, the same assumptions

hold for the estimated $\widehat{\mathbf{F}}$ in probability. We state these results formally in Lemmas 6.2 and 6.3 in Section 6.

5. Analysis Under Sample Incoherence Matrix Assumptions. In this section, we provide a proof overview of Proposition 5.1 and the key technical contribution of our paper. All the detailed statements and proofs of the Proposition 5.1 and its relevant Lemmas are relegated to the supplementary material for the conciseness.

5.1. Statement of Proposition. We establish the signed-support consistency of ℓ_1-PsLS estimator when the assumptions are directly imposed on the estimated feature matrix $\widehat{\mathbf{F}}$, instead on the ground-truth feature matrix $\mathbf{F}$. More specifically, we assume that there exist some constants $\mu \in (0, 1]$ and $C_{\min} > 0$, such that the followings hold:

$$\text{(A3)} \qquad \mathbb{P}\left[\left\| \widehat{\mathcal{Q}}_N \right\|_\infty \leq 1 - \mu \right] \geq P_\mu \quad \text{and} \quad \Lambda_{\min}\Big(\frac{1}{NM} \widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}} \Big) \geq C_{\min} \quad \text{almost surely} .$$

Here, $P_\mu \in [0, 1]$ denotes some probability that $\widehat{\mathcal{Q}}_N$ satisfies the incoherence assumption. Equipped with this assumption, we have the following proposition:

Proposition 5.1. *Given the observed data set $\mathcal{D}$, where the spatial resolution is related to the temporal resolution via $M = \Theta(N^{\frac{2P_{\max}+5}{7}})$, we take the bandwidths of the kernels in (3.2) and (3.3) as $h_N = \Theta(N^{-\frac{1}{7}})$, $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. Under the assumptions in (A3) imposed on the estimated feature matrix $\widehat{\mathbf{F}}$, suppose that the sequence of regularization hyper-parameters $\{\lambda_N\}$ satisfies $\lambda_N = \Omega\Big(\frac{\sqrt{K} \ln N}{\mu N^{2/7 - c}} \Big)$ for some constant $0 < c < \frac{2}{7}$ independent of N. Then, the following properties hold :*

(i) With probability greater than $P_\mu - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}} \exp\big(-\frac{1}{6} N^c \big) \right) \to P_\mu$ as $N \to \infty$, the ℓ_1-PsLS method (3.7) has a unique minimizer $\widehat{\boldsymbol{\beta}}^\lambda \in \mathbb{R}^K$ with its support contained within the true support, that is $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$.

(ii) With probability greater than $1 - \mathcal{O}\left(N^{\frac{2P_{\max}+5}{7}} \exp\big(-\frac{1}{6} N^c \big) \right) \to 1$ as $N \to \infty$, $\widehat{\boldsymbol{\beta}}^\lambda$ satisfies the ℓ_∞ bound:

$$\text{(5.1)} \qquad \left\| \widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}} - \boldsymbol{\beta}^*_{\mathcal{S}} \right\|_\infty \leq K^{3/2} C_{\min} \left(o_N(1) + \lambda_N \right) .$$

(iii) Additionally, if the minimum value of model parameter supported on $\mathcal{S}$ is greater than the upper-bound of (5.1), that is $\min_{1 \leq i \leq s} |(\boldsymbol{\beta}^*_{\mathcal{S}})_i| > K^{3/2} C_{\min} \left(o_N(1) + \lambda_N \right)$, then $\widehat{\boldsymbol{\beta}}^\lambda$ has a correct signed-support. (i.e., $\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}^\lambda) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)$)

We remark that the first item (i) in Proposition 5.1 holds with probability $P_\mu \leq 1$ asymptotically, while the second item (ii) holds with probability 1 asymptotically. They are not contradictory, since (i) describes the support recovery of the coefficient vector over all indices, whereas (ii) focuses on the estimation errors on entries within the true support $\mathcal{S}$. Technically speaking, proof of (i) is involved with mutual incoherence condition in (A3), whereas (ii) is involved with minimum-eigen value condition on $\widehat{\mathbf{F}}$ in (A3).

5.2. Proof Overview of Proposition 5.1. Readers can find the proof of (5.1) in the supplementary material SM 3.6. Here, we focus on providing the high-level idea on the proof of (i) of Propostion 5.1. The most important ingredient for the success of PDW construction is to establish the *strict dual feasibility* of the dual vector $\widehat{\mathbf{z}}$, when $\widehat{\mathbf{z}} \in \partial\|\widehat{\boldsymbol{\beta}}^{\lambda}\|_1$, where $\partial\|\widehat{\boldsymbol{\beta}}^{\lambda}\|_1$ is a sub-differential set of $\|\cdot\|_1$ evaluated at $\widehat{\boldsymbol{\beta}}^{\lambda}$. In other words, we need to ensure that $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1$ with high probability. (See supplementary material SM1.) Through Karush–Kuhn–Tucker (KKT) condition of the optimal pair $(\widehat{\boldsymbol{\beta}}^{\lambda}, \widehat{\mathbf{z}})$ of (3.7) and settings of PDW construction, we can explicitly derive the expression of the dual vector $\widehat{\mathbf{z}}$ supported on the complement of the support set $\mathcal{S}$ as follows:

$$\widehat{\mathbf{z}}_{\mathcal{S}^c} = \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \widehat{\mathbf{F}}_{\mathcal{S}} (\widehat{\mathbf{F}}_{\mathcal{S}}^T \widehat{\mathbf{F}}_{\mathcal{S}})^{-1} \widehat{\mathbf{z}}_{\mathcal{S}} + \underbrace{\frac{1}{\lambda_N MN} \widehat{\mathbf{F}}_{\mathcal{S}^c}^T \boldsymbol{\Pi}_{\mathcal{S}^\perp} (\Delta \mathbf{u}_t - \Delta \mathbf{F}_{\mathcal{S}} \boldsymbol{\beta}_{\mathcal{S}}^*)}_{:=\tilde{Z}_{\mathcal{S}^c}}, \tag{5.2}$$

where $\boldsymbol{\Pi}_{\mathcal{S}^\perp}$ is an orthogonal projection operator on the column space of $\widehat{\mathbf{F}}_{\mathcal{S}}$. By the mutual incoherence condition in (A3), the first term of the right-hand side in (5.2) is upper-bounded by $1-\mu$ for some $\mu \in (0,1]$, with some probability $P_\mu \in [0,1]$. The remaining task is to control the tail probability of $\tilde{Z}_j$ for $j \in \mathcal{S}^c$: that is to ensure $\mathbb{P}\big[\max_{j\in\mathcal{S}^c} |\tilde{Z}_j| \geq \mu\big] \to 0$ with some exponential decay rate. With the help of Lemma SM3.1 in the supplementary material, controlling the probability $\mathbb{P}\big[\|\tilde{Z}_{\mathcal{S}^c}\|_\infty \geq \mu\big]$ reduces to controlling $\mathbb{P}\big[\|\Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* - \Delta\mathbf{u}_t\|_\infty \geq \mu\frac{\lambda_N}{\sqrt{K}}\big]$. Controlling the bound on $\mathbb{P}\big[\|\boldsymbol{\tau}\|_\infty \geq \varepsilon\big]$ for some $\varepsilon > 0$ is challenging, since the exact form of the residual distribution $\boldsymbol{\tau}$ is unknown. (Note that $\boldsymbol{\tau} = \Delta\mathbf{F}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}^* - \Delta\mathbf{u}_t$ since $\mathbf{u}_t = \mathbf{F}\boldsymbol{\beta}^*$.)

We circumvent this difficulty by using the following inequality: for some thresholds $\varepsilon_N > 0$ and $\varepsilon_M > 0$, both of which go to 0 as N and M tends to ∞, we have,

$$\begin{aligned}
&\mathbb{P}\big[\,\|\boldsymbol{\tau}\|_\infty \geq \varepsilon_N + \varepsilon_M\big] \\
&\leq \mathbb{P}\Big[\max_{0\leq i\leq M-1} \sup_{t\in[0,T_{\max})} |\Delta u_t(X_i,t)| \geq \varepsilon_N\Big] + \mathbb{P}\Big[\max_{\substack{1\leq k\leq s\\ 0\leq n\leq N-1}} \sup_{x\in[0,X_{\max})} |\Delta F_k(x,t_n)| \geq \frac{\varepsilon_M}{s\|\boldsymbol{\beta}^*\|_\infty}\Big] \\
&\leq M\cdot\mathbb{P}\Big[\sup_{t\in[0,T_{\max})} |\Delta u_t(X_i,t)| \geq \varepsilon_N\Big] + sN\cdot\mathbb{P}\Big[\sup_{x\in[0,X_{\max})} |\Delta F_k(x,t_n)| \geq \frac{\varepsilon_M}{s\|\boldsymbol{\beta}^*\|_\infty}\Big].
\end{aligned}$$

The above inequality naturally leads us to study the uniform convergence of Local-Polynomial estimator to its ground-truth function of interest. Say, for sufficiently large enough grid size of temporal dimension N, for some $\varepsilon_N \geq 0$ that is h_N-dependent threshold and $X_i \in [0, X_{\max})$, we will achieve

$$\mathbb{P}\Big[\sup_{t\in[0,T_{\max})} |\widehat{\mathbf{u}}_t(X_i,t) - \mathbf{u}_t(X_i,t)| > \varepsilon_N\Big] \to 0, \tag{5.3}$$

with an exponential decay rate. As for obtaining the exponential decay rate in (5.3), we defer the detailed explanation with some intuitions in the following Subsection. It turns out that thresholds ε_N and ε_M are functions of bandwidth parameters h_N and w_M in (3.4) and (3.5). We choose correct orders of h_N and w_M so that we can ensure that the thresholds ε_N and ε_M go to zero. Then, with the proper choice on the order of λ_N together with $\mathbb{P}\big[\|\boldsymbol{\tau}\|_\infty \geq \mu\frac{\lambda_N}{\sqrt{K}}\big]$, we conclude the proof.

5.3. Technical Contribution. Several researchers have tried to achieve uniform convergence of Local-Polynomial or kernel smoothing estimators in almost sure sense. See the works [45] and [46]. However, to the best of the authors' knowledge, uniform convergence of Local-Polynomial estimator with an explicit decaying probability rate has not been studied in the literature. We provide it as a technical contribution of the present paper. Readers can find the exact statements of these results for the estimators $\widehat{\mathbf{u}}_t$ and $\widehat{\partial_x^p u}$ for $0 \leq p \leq P_{\max}$ in the supplementary material stated as Lemma SM3.2 and Lemma **SM3.3**, respectively.

Here, we provide a high-level idea of the proof of Lemma SM3.2. First, we observe that the higher-order Local-Polynomial smoothing is asymptotically equivalent to higher-order kernel smoothing through equivalent kernel theory [8]. See (3.4) and (3.5) for their equivalences in mathematical form with kernel smoothing estimators. Second, we employ the truncation idea in [43] on the Local-Polynomial estimator and decompose $\widehat{\mathbf{u}}_t(X_i,t) - \mathbf{u}_t(X_i,t)$ into three parts as follows:

$$\widehat{\mathbf{u}}_t - \mathbf{u}_t = \underbrace{\left(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N} - \mathbb{E}\big(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}\big)\right)}_{\text{Asymptotic deviation of truncation error}} + \underbrace{\left(\widehat{\mathbf{u}}_t^{B'_N} - \mathbb{E}\widehat{\mathbf{u}}_t^{B'_N}\right)}_{\substack{\text{Asymptotic deviation of} \\ \text{truncated estimator}}} + \underbrace{\left(\mathbb{E}\widehat{\mathbf{u}}_t - \mathbf{u}_t\right)}_{\text{Asymptotic bias}},$$

where B'_N is some increasing sequence in N, and $\widehat{\mathbf{u}}_t^{B'_N}$ denotes the truncated Local-Polynomial estimator of $\mathbf{u}_t$. We control the *sup* over $t \in [0, T_{\max})$ on each of these three components. The last component, *Asymptotic bias* of $\widehat{\mathbf{u}}_t$ can be obtained through the classical result from [8, 9]. The exponential decay rate comes from the first two components as follows:

1. *Asymptotic deviation of truncation error* can be decomposed into two parts. The first part, which is $\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}$, can be easily controlled via Chernoff bound of Gaussian random variable. by using the definition of truncated estimator $\widehat{\mathbf{u}}_t^{B'_N}$. The second part, which is the expected difference $\mathbb{E}\big(\widehat{\mathbf{u}}_t - \widehat{\mathbf{u}}_t^{B'_N}\big)$, can be bounded by some deterministic function of B'_N and h_N using the similar arguments in Proposition 1 of [43].

2. *Asymptotic deviation of truncated estimator* is decomposed into two components as well: (i) Brownian bridge and (ii) difference between some two-dimensional empirical process and the Brownian bridge. (i) can be controlled via uniform convergence of Gaussian Process using the arguments similar to [47], together with simple Markov inequality. (ii) can be controlled via Tusnády's strong uniform approximation theory [43, 44], stating that the two-dimensional empirical process can be well approximated by a certain solution path of two-dimensional Brownian bridge.

Same ideas can be employed for the uniform convergence of $\widehat{(\partial_x^p u)_i^n}$ to $(\partial_x^p u)_i^n$ and of $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}$ to $(\partial_x^p u)_i^n(\partial_x^q u)_i^n$ for $0 \leq p, q \leq P_{\max}$.

6. Uniform Convergence of Sample Incoherence Matrix. In this section, we provide two Lemmas 6.2 and 6.3 that can complete the proof of Theorem 4.1. Here, the minimum-eigenvalue and incoherence assumptions are imposed on the ground-truth feature matrix $\mathbf{F}$, instead on the estimated feature matrix $\widehat{\mathbf{F}}$. See (A1) and (A2). That is, there exist $C_{\min} > 0$

and $\mu \in (0,1]$ such that the followings hold for the unknown support set $\mathcal{S}$:

$$\Lambda_{\min}\Big(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{T}\mathbf{F}_{\mathcal{S}}\Big) \geq C_{\min} \quad \text{and} \quad \big\|\mathcal{Q}^{*}\big\|_{\infty} \leq 1-\mu.$$

Equipped with the above assumptions, we can formally show that success probability of the sample incoherence condition P_μ in (A3) tends to 1 as $N \to \infty$.

The key step of the proofs in the following Lemmas is to control the tail probability of difference between inner-product of two arbitrary columns of $\widehat{\mathbf{F}}$ and inner-product of the two corresponding columns of ground-truth $\mathbf{F}$. This problem is challenging even if the exact distribution of any entries of $\widehat{\mathbf{F}}$ is known, since the distribution of $\sum_{k=1}^{NM}\widehat{F}_{ki}\widehat{F}_{kj}$ needs to be derived. In order to circumvent this problem, we take the advantage of the uniform convergence result of $\widehat{(\partial_x^p u)_i^n}$ for any $0 \leq p \leq P_{\max}$ proved in Lemma **SM3.3**. Additionally, we need the uniform convergence results of $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}$, $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}\widehat{(\partial_x^k u)_i^n}$, and $\widehat{(\partial_x^p u)_i^n}\widehat{(\partial_x^q u)_i^n}\widehat{(\partial_x^k u)_i^n}\widehat{(\partial_x^l u)_i^n}$ for $0 \leq p,q,k,l \leq P_{\max}$. These convergence results are explicitly stated as Corollaries **SM3.7**, **SM4.1**, and **SM4.2**, with proofs in the supplementary material.

Equipped with the convergence results, we introduce a following Lemma stating that the distance between the matrices $\widehat{\mathbf{F}}_{\mathcal{S}^c}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}$ and $\mathbf{F}_{\mathcal{S}^c}^{\top}\mathbf{F}_{\mathcal{S}}$ are close enough under operator norm for large enough grid sizes.

Lemma 6.1. *Let $\varepsilon_M^{*}, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}$ be the thresholds defined in **SM3.3**, **SM3.7**, **SM4.1**, and **SM4.2**. Then for any $\varepsilon_M^{max'}$ such that*

$$\varepsilon_M^{max'} > \sqrt{s(K-s)}\max\left\{\varepsilon_M^{*}, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}\right\},$$

then, for $0 < c < \frac{2}{7}$, and for sufficiently large enough N, we have

$$\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{max'}\right] \leq \mathcal{O}\Big(N\exp\big(-\frac{1}{6}N^{c}\big)\Big).$$

Now, we are ready to prove our main claims, Lemmas 6.2 and 6.3. We first state and prove the Lemma 6.2 asserting that if there exists $C_{\min} > 0$ such that the minimum eigen-value condition holds for $\mathbf{F}_{\mathcal{S}}$, then the sample minimum eigen-value condition holds with probability converging to 1 with an exponential decay rate.

Lemma 6.2. *Suppose that the assumption (A1) holds with some constant $C_{min} > 0$ and $0 < c < \frac{2}{7}$, then with probability at least $1 - \mathcal{O}(N\exp(-\frac{1}{6}N^{c})) \to 1$ as $N \to \infty$, we have,*

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

Proof. Observe that we can write:

$$\begin{aligned}\Lambda_{\min}\left(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right) &:= \frac{1}{NM}\min_{\|x\|_2=1} x^{\top}\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right)x \\ &= \frac{1}{NM}\min_{\|x\|_2=1}\left\{x^{\top}\left(\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)x + x^{\top}\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)x\right\} \\ &\leq \frac{1}{NM}\left\{y^{\top}\left(\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)y + y^{\top}\left(\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)y\right\}\end{aligned}$$

where $y \in \mathbb{R}^K$ is a unit-norm minimal eigen-vector of $\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}$. Therefore, we can write,

$$\begin{aligned}\Lambda_{\min}\left(\frac{1}{NM}\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right) &\geq \Lambda_{\min}\left(\frac{1}{NM}\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right) - \frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2 \\ &\geq C_{\min} - \frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_2 .\end{aligned}$$

By using a similar argument used in Lemma 6.1, we can prove $\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_2 \to 0$ with high-probability as $N \to \infty$. For any $\varepsilon_M^{\max}$ such that,

$$\varepsilon_M^{\max} > s\max\left\{\varepsilon_M^{*}, \varepsilon_M^{**}, \varepsilon_M^{***}, \varepsilon_M^{****}\right\}, \tag{6.1}$$

Then, we can bound the probability as follows:

$$\begin{aligned}&\mathbb{P}\left[\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_2 > \varepsilon_M^{\max}\right] \\ &\quad\leq \mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_{\mathrm{F}} > NM\varepsilon_M^{\max}\right] \leq \mathbb{P}\left[\left\|\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}}^{\top}\mathbf{F}_{\mathcal{S}}\right\|_{\infty,\infty} > NM\frac{\varepsilon_M^{\max}}{s}\right] \\ &\quad\leq \mathbb{P}\left[\max_{n=0,\ldots,N-1}\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}_i}(x,t_n)\widehat{\mathbf{F}_j}(x,t_n) - \mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\max}}{s}\right] \\ &\quad\leq \sum_{n=0}^{N-1}\mathbb{P}\left[\sup_{x\in[0,X_{\max})}\left|\widehat{\mathbf{F}_i}(x,t_n)\widehat{\mathbf{F}_j}(x,t_n) - \mathbf{F}_i(x,t_n)\mathbf{F}_j(x,t_n)\right| > \frac{\varepsilon_M^{\max}}{s}\right] \\ &\quad\leq \mathcal{O}\left(N\exp\left(-\frac{1}{6}N^c\right)\right).\end{aligned}$$

■

With the help of Lemma 6.2, we can show that the sample incoherence condition holds with high probability, given that there exists $\mu \in (0,1]$ for the ground-truth version of (A2).

Lemma 6.3. *Suppose that the assumption (A2) holds with some constant $\mu \in (0,1]$ and $0 < c < \frac{2}{7}$, then with probability at least $1 - \mathcal{O}(N\exp(-\frac{1}{6}N^c)) \to 1$ as $N \to \infty$, we have,*

$$\left\|\widehat{\mathcal{Q}}_N\right\|_{\infty} \leq 1 - \mu .$$

Proof. Motviated from [18], we begin the proof by decomposing the matrix $\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$ into four parts:

$$\begin{aligned}\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} &= \underbrace{\mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right)}_{:=\mathbf{T_1}} + \underbrace{\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}}_{:=\mathbf{T_2}} \\ &+ \underbrace{\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1} - \left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right)}_{:=\mathbf{T_3}} \\ &+ \underbrace{\left(\mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}}_{:=\mathbf{T_4}}.\end{aligned}$$

Since we know $|||\mathbf{T_4}|||_\infty \leq 1-\mu$ for some $\mu \in (0,1]$, the decomposition reduces the proof showing $|||\mathbf{T_i}|||_\infty \to \mathbf{0}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ for $i=1,2,3$.

1. Control of $\mathbf{T_1}$: Observe that we can re-factorize $\mathbf{T_1}$ as follows:

$$\mathbf{T_1} = \left(\mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}.$$

Then, by taking the advantage of sub-multiplicative property $|||AB|||_\infty \leq |||A|||_\infty |||B|||_\infty$ and the fact $|||\mathbf{T_4}|||_\infty \leq 1-\mu$ and $|||C|||_\infty \leq \sqrt{N}|||C|||_2$ for $C \in \mathbb{R}^{M\times N}$, we can bound $|||\mathbf{T_1}|||_\infty$ as follows:

$$\begin{aligned}|||\mathbf{T_1}|||_\infty &\leq \left|\left|\left|\left(\mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right)\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_\infty \left|\left|\left|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right|\right|\right|_\infty \left|\left|\left|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_\infty \\ &\leq s(1-\mu)\left(\frac{1}{NM}\left|\left|\left|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right|\right|\right|_2\right)\left(NM\left|\left|\left|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_2\right) \\ &\leq \frac{s(1-\mu)}{C_{\min}}\left(\frac{1}{NM}\left|\left|\left|\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}} - \widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right|\right|\right|_2\right).\end{aligned}$$

Note that we use $|||\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}|||_2 \leq \frac{1}{NMC_{min}}$ with probability $1-\mathcal{O}(N\exp(-\frac{1}{6}N^c))$ in the last inequality from Lemma 6.1.

2. Control of $\mathbf{T_2}$: With similar techniques employed for controlling $|||\mathbf{T_1}|||_\infty$, we can bound $|||\mathbf{T_2}|||_\infty$ as follows:

$$\begin{aligned}|||\mathbf{T_2}|||_\infty &\leq \left|\left|\left|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right|\right|\right|_\infty \left|\left|\left|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_\infty \\ &\leq s\left|\left|\left|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right|\right|\right|_2 \left|\left|\left|\left(\mathbf{F}_{\mathcal{S}}^\top \mathbf{F}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_2 \\ &= s\left(\frac{1}{NM}\left|\left|\left|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right|\right|\right|_2\right)\left(NM\left|\left|\left|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top \widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right|\right|\right|_2\right) \\ &\leq \frac{s}{C_{\min}}\left(\frac{1}{NM}\left|\left|\left|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top \widehat{\mathbf{F}}_{\mathcal{S}} - \mathbf{F}_{\mathcal{S}^c}^\top \mathbf{F}_{\mathcal{S}}\right|\right|\right|_2\right).\end{aligned}$$

3. Control of $\mathbf{T_3}$: To bound $\left\|\mathbf{T_3}\right\|_\infty$, we re-factorize the second argument of product in $\mathbf{T_3}$:

$$\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}=\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)-\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}$$

With the factorization, we bound $\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\|_\infty$ by using sub-multiplicative property and the fact $\|C\|_\infty\le\sqrt{N}\|C\|_2$ for any $C\in\mathbb{R}^{M\times N}$ again:

$$\begin{aligned}\left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_\infty&=\left\|\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)-\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_\infty\\&\le\sqrt{s}\left\|\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\left[\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)-\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_2\\&\le\sqrt{s}\left\|\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_2\left\|\left[\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)-\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)\right]\right\|_2\left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\right\|_2\\&\le\frac{\sqrt{s}}{NMC_{\min}^2}\left(\frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}-\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\right).\end{aligned}\tag{6.2}$$

In the last inequality, we use the result of Lemma 6.1. Now we can bound $\left\|\mathbf{T_3}\right\|_\infty$ as follows:

$$\begin{aligned}\left\|\mathbf{T_3}\right\|_\infty&=\left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\right)\left(\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\right)\right\|_\infty\\&\le\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\right\|_\infty\left\|\left(\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}-\left(\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}\right)^{-1}\right\|_\infty\\&\le\frac{s}{C_{\min}}\left(\frac{1}{NM}\left\|\widehat{\mathbf{F}}_{\mathcal{S}^c}^\top\widehat{\mathbf{F}}_{\mathcal{S}}-\mathbf{F}_{\mathcal{S}^c}^\top\mathbf{F}_{\mathcal{S}}\right\|_2\right)\left(\frac{1}{NM}\left\|\mathbf{F}_{\mathcal{S}}^\top\mathbf{F}_{\mathcal{S}}-\widehat{\mathbf{F}}_{\mathcal{S}}^\top\widehat{\mathbf{F}}_{\mathcal{S}}\right\|_2\right),\end{aligned}$$

where in the last inequality, we use (6.2) and $\|C\|_\infty\le\sqrt{N}\|C\|_2$ for any $C\in\mathbb{R}^{M\times N}$. Take $\varepsilon_M^{\max''}$ such that,

$$\varepsilon_M^{\max''}>\max\left\{\frac{C_{\min}}{s(1-\mu)}\varepsilon_M^{\max},\frac{C_{\min}}{s}\varepsilon_M^{\max'}\right\},$$

with $\varepsilon_M^{\max}$ in (6.1) and with $\varepsilon_M^{\max'}$ in Lemma 6.1, respectively. Then, we have

$$\mathbb{P}\left[\forall i=1,2,3:\left\|\mathbf{T_i}\right\|_\infty>\varepsilon_M^{\max''}\right]\le\mathcal{O}\left(N\exp\left(-\frac{1}{6}N^c\right)\right).$$ ■

Verification of Lemma 6.3 automatically leads to the complete proof of Theorem 4.1, together with Proposition 5.1. Therefore, as long as the two assumptions (A1) and (A2) hold for $\mathbf{F}$, with sufficiently fine-grained grid points over the function $u(X,t)$, ℓ_1-PsLS can always find the correct signed-support of the given PDE model, with the minimum absolute value of $\boldsymbol{\beta}_{\mathcal{S}}^*$ not too close to zero.

7. Numerical Experiments.

In the first subsection, two PDE models and data-generating processes of respective models are introduced. In the next subsection, we verify the main statements of the Theorem 4.1 through numerical experiments over the PDE models described in Subsection 7.1. The impact of $\boldsymbol{\beta}_{\min}^*$-condition in the signed-support recovery of ℓ_1-PsLS is numerically explored in subsection 7.3.

7.1. Experimental Setting. In this subsection, we provide detailed descriptions on (i) two popular PDE models that we are going to work on throughout the Section 7, and on (ii) how to generate the data from respective models, and (iii) how to design the regression problem for the experiments to be presented.

7.1.1. Model Specification and Data Generation. **Viscous Burgers' equation** is a fundamental second-order semilinear PDE which is frequently employed to model physical phenomena in fluid dynamics [48] and nonlinear acoustic in dissipative media [49]. Its general form is

$$u_t = -uu_x + \nu u_{xx}$$

where $\nu > 0$ is the diffusion coefficient which characterizes physical quantities such as viscosity of fluid. Specifically, when $\nu = 0$, it becomes an inviscid Burgers' equation, which is a conservative system that can form shock waves. Here we consider the following viscous Burgers' equation:

$$\begin{aligned} &u_t = -uu_x + \nu u_{xx} \ , \ \ 0 < x < 1, 0 < t < 0.1 \\ &u(x,0) = \sin^2(2\pi x) + \cos^3(3\pi x) \ , \ \ 0 \le x \le 1 \ , \ \ u(0,t) = u(1,t) \ , \ \ 0 \le t \le 0.1. \end{aligned} \tag{7.1}$$

Korteweg–de Vries equation is well known for its solution that demonstrates the phenomenon of superposition of nonlinear waves [50], and for modeling fluid dynamics of shallow water surfaces in long and narrow channels [51]. Its dimensionless form is given as

$$u_t + u_{xxx} + 6uu_x = 0 \ . \tag{7.2}$$

In this Section, we consider the form of (7.2), whose initial solution is as follows:

$$\begin{aligned} &u(x,0) = 3.5\sin^3(4\pi x) + 1.5\exp\big(-\sin(2\pi x)(1-x)\big) \ , \\ &0 \le x \le 1 \ , \ \ u(0,t) = u(1,t) \ , \ \ 0 \le t \le 0.1. \end{aligned}$$

Data Generation For N-size sampling in the temporal dimension, by Theorem 4.1, we take $M = \lfloor N^{(2\times P_{\max}+5)/7} \rfloor$ sample size in the space dimension. We numerically solve Viscous Burgers' equation (7.1) by the Lax-Wendroff scheme on a grid with interval width $\delta t = 0.1/(100N)$ in temporal and $\delta x = 1/M$ in space, then we downsampled the data in the temporal dimension by a factor of 100; thus the resulted clean data is distributed over a grid with N nodes in time and M nodes in space. Lastly, we added i.i.d. Gaussian noise with standard deviation $\sigma = 0.25$ to the data. i.e., $\nu_i^n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 0.25^2)$. As for solving the KdV equation (7.2), the same approaches with Viscous Burger's equation are applied, with i.i.d. Gaussian noises with standard deviation $\sigma = 0.025$.

7.1.2. Constructions of Regression Problems. We employ the Local-Polynomial smoothing for estimating $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ as described in Subsection 3.2. Regarding a choice of kernel for constructing $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$, we use the Epanechnikov kernel defined by:

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

Table 1

Specific choices of the constants in the order of $h_N = \Theta(N^{-\frac{1}{7}})$ and $w_M = \Theta(M^{-\frac{1}{7}})$ for the experiments on Viscous Burgers equation and KdV equation are presented.

	w_M	h_N
Viscous Burgers	$0.75M^{-\frac{1}{7}}$	$0.25N^{-\frac{1}{7}}$
KdV	$0.1M^{-\frac{1}{7}}$	$0.01N^{-\frac{1}{7}}$

where $(\cdot)_+ := \max(0, \cdot)$. Bandwidth parameters h_N and w_M in (3.2) and (3.3) are chosen in the order of $h_N = \Theta(N^{-\frac{1}{7}})$ and $w_M = \Theta(M^{-\frac{1}{7}})$, respectively. As displayed in Table 1, for the experiments presented in this Section, we choose specific constant factors in the order expressions of h_N and w_M for Viscous Burgers equation and KdV equation. Regarding more detailed issues on the choices of these constants, readers can refer to Section 8. It is also worth noting that we do not use (3.4) and (3.5) as solutions of the optimization problems (3.2) and (3.3) for the experiments, since the expressions in (3.4) and (3.5) are derived in asymptotic settings. For the reader's convenience, We provide the closed form solutions of (3.4) and (3.5) in supplementary material SM2.

For Viscous Burgers' equation with noisy data, Local-Polynomial fitting with $P_{\max} = 2$ is applied to construct $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$. Our goal is to identify the fifth and the sixth coefficients, $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_6$, of a following linear measurement via the proposed ℓ_1-PsLS model (3.7):

$$\widehat{\mathbf{u}}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\widehat{\mathbf{u}} + \boldsymbol{\beta}_2\widehat{\mathbf{u}}^2 + \boldsymbol{\beta}_3\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_4\widehat{\mathbf{u}}_x^2 + \boldsymbol{\beta}_5\widehat{\mathbf{u}}\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_6\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_7\widehat{\mathbf{u}}_{xx}^2 + \boldsymbol{\beta}_8\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_9\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xx}.$$

For KdV equation, after generating the data-points, $\widehat{\mathbf{u}}_t$ and $\widehat{\mathbf{F}}$ are fitted through Local-Polynomial with $P_{max} = 3$. We want ℓ_1-PsLS to select $\boldsymbol{\beta}_5$ and $\boldsymbol{\beta}_{10}$ as non-zero coefficients in a following linear measurement:

$$\begin{aligned}\widehat{\mathbf{u}}_t = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1\widehat{\mathbf{u}} + \boldsymbol{\beta}_2\widehat{\mathbf{u}}^2 + \boldsymbol{\beta}_3\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_4\widehat{\mathbf{u}}_x^2 + \boldsymbol{\beta}_5\widehat{\mathbf{u}}\widehat{\mathbf{u}}_x + \boldsymbol{\beta}_6\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_7\widehat{\mathbf{u}}_{xx}^2 + \boldsymbol{\beta}_8\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xx} + \boldsymbol{\beta}_9\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xx} \\ + \boldsymbol{\beta}_{10}\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{11}\widehat{\mathbf{u}}_{xxx}^2 + \boldsymbol{\beta}_{12}\widehat{\mathbf{u}}_x\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{13}\widehat{\mathbf{u}}_{xx}\widehat{\mathbf{u}}_{xxx} + \boldsymbol{\beta}_{14}\widehat{\mathbf{u}}\widehat{\mathbf{u}}_{xxx}.\end{aligned}$$

7.2. Numerical Verifications of Main Statements. In this subsection, we design an experiment to numerically verify following two main statements of this paper. [2]

1. *Under assumptions (A1) and (A2), and with large enough data points, there exist some $\lambda_N \geq 0$ such that ℓ_1-PsLS model (3.7) recovers a signed-support $\left(\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}^\lambda) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)\right)$ of an unique PDE that admits the underlying function as a solution in probability.*
2. *Given assumption (A2) for some $\mu \in (0, 1]$, sampled incoherence parameter μ' converges to ground-truth μ in probability with large enough data points.*

The experiment is conducted over two PDE models, **Viscous Burgers' equation** and **KdV equation** introduced in Subsection 7.1. We generate the data by setting $\nu = 0.03$ in (7.1). In Figure 1, the probability of signed-support recovery $\mathbb{P}[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)]$ versus the grid size of temporal dimension N, and $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty$ versus N are recorded on the same plot for

[2]Results provided in Subsections 7.2 and 7.3 can be reproduced via MATLAB codes in *https://github.com/namjoonsuh/PDE-identification.*

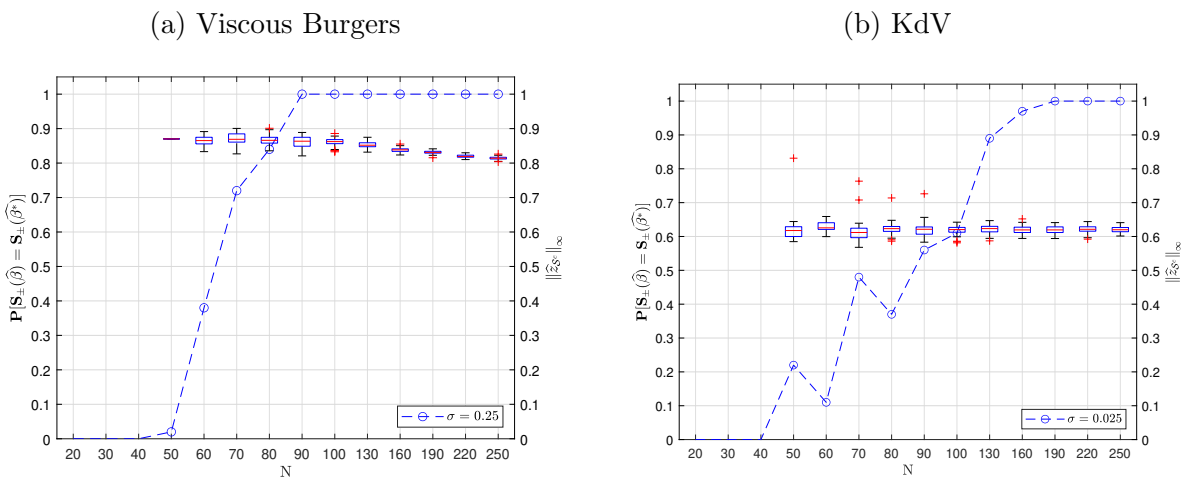

Figure 1. *Probability of signed-support recovery* $\mathbb{P}[\mathbb{S}_\pm(\widehat{\beta}) = \mathbb{S}_\pm(\beta^*)]$ *versus the grid size of temporal dimension* N*, and* $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty$ *versus* N *are recorded on the same plot for Viscous Burger's equation in panel (a) and for KdV equation in panel (b), respectively.*

respective models. Each point on each curve, which represents $\mathbb{P}[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)]$, in (a) and (b) corresponds to the average over 100 trials. For each iteration, the hyper-parameter λ_N is chosen in an "optimal" way: we used the value yielding the correct number of nonzero coefficient. With the chosen λ_N, $\widehat{\mathbf{z}}_{\mathcal{S}^c}$ is calculated as given in (5.2). Note that (5.2) can be calculated only when the ℓ_1-PsLS finds λ_N that gives the minimizer of (3.7) $\widehat{\boldsymbol{\beta}}^\lambda$ such that $\widehat{\boldsymbol{\beta}}^\lambda_{\mathcal{S}^c} = 0$ and $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$. For this reason, boxplots of $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty$ in (a) and (b) are drawn from the point when ℓ_1-PsLS starts to find such λ_N. For both models, $\mathbb{P}[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)]$ goes to 1, as we observe more data points on finer grid. Furthermore, it is worth noting that the *strict dual feasibility* condition (i.e., $\|\widehat{\mathbf{z}}_{\mathcal{S}^c}\|_\infty < 1$) holds for both cases. In Figure 2, boxplots of $|||\widehat{\mathcal{Q}}_N|||_\infty$ versus N are displayed for Viscous Burgers' equation and kDV equation respectively. A dotted horizontal line in each panel represents $1 - \mu$ calculated from the ground-truth feature matrix $\mathbf{F}$. Notice that as the number of observed data gets larger, the sampled incoherence parameter goes below the dotted lines for both models.

7.3. Impact of $\beta^*_{\min}$ in Signed-Support Recovery of ℓ_1-PsLS. Theorem 4.1 states that as long as $\beta^*_{\min} := \min_{i\in\mathcal{S}} |\beta^*_i|$ is beyond certain threshold, ℓ_1-PsLS is signed-support recovery consistent. In this subsection, we design an experiment to numerically confirm this claim. The experiment is performed over Viscous Burgers' equation by varying the coefficient ν in (7.1) : we set $\nu = 0.03, 0.02, 0.01, 0.005$ The Figure 3 (a) displays the curves representing $\mathbb{P}[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)]$ versus N for each of the four cases. Each point on each curve represents the average over 100 trials. The Figure 3 (b) exhibits the range of λ_N for which ℓ_1-PsLS finds the support of $\widehat{\boldsymbol{\beta}}^\lambda$ that is contained within the true support, when ν is set as 0.005. More specifically, boxplots in (b) record the range of λ_N that picks $\widehat{\mathbf{u}}_{xx}$ as the selected argument. In (a), we can check that, as the magnitude of $\min_{i\in\mathcal{S}} |\boldsymbol{\beta}^*_i|$ decreases from 0.03 to 0.01, ℓ_1-PsLS requires more data-points for the signed-support recovery, and when $\min_{i\in\mathcal{S}} |\boldsymbol{\beta}^*_i|$ drops

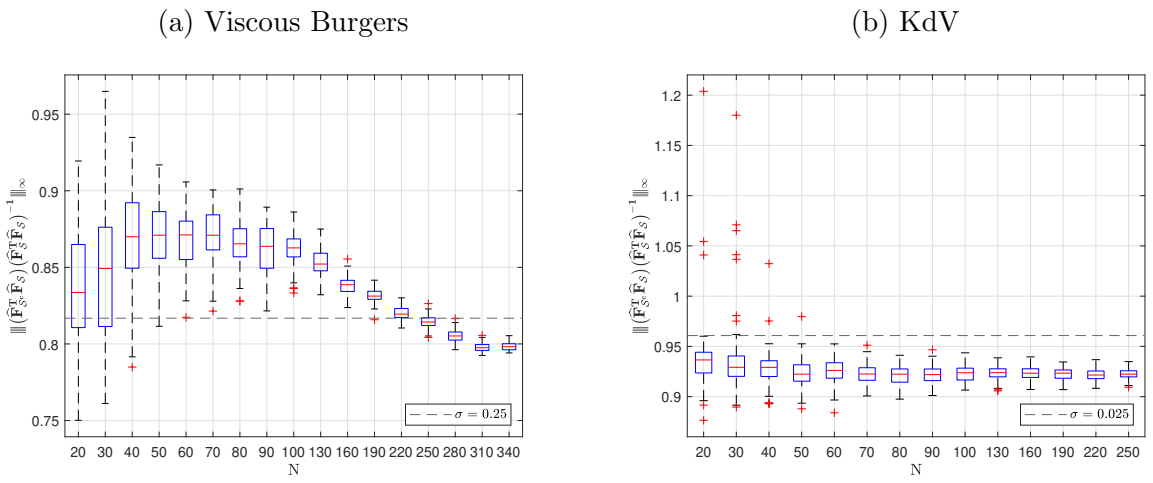

Figure 2. *Boxplots of* $\|\widehat{\mathcal{Q}}_N\|_\infty$ *versus* N *are displayed for Viscous Burgers' equation in panel (a) and KdV equation in panel (b), respectively.*

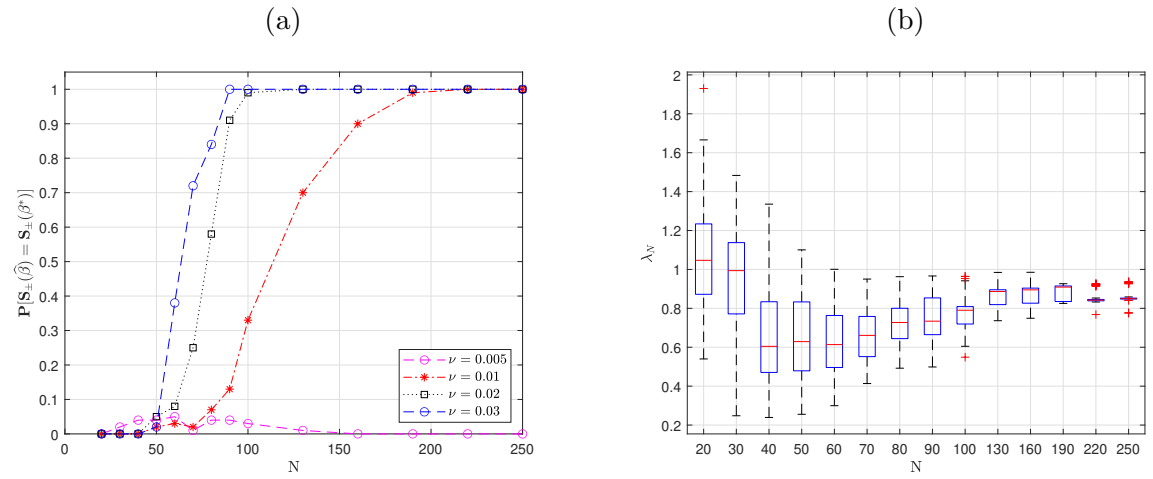

Figure 3. *Left panel (a) displays the curves representing* $\mathbb{P}[\mathbb{S}_\pm(\widehat{\boldsymbol{\beta}}) = \mathbb{S}_\pm(\boldsymbol{\beta}^*)]$ *versus* N*, when* $\nu = 0.03, 0.02, 0.01, 0.005$*. Right panel (b) exhibits the range of* λ_N *for which* ℓ_1*-PsLS gives the solution* $\widehat{\boldsymbol{\beta}}^\lambda$ *such that* $\mathcal{S}(\widehat{\boldsymbol{\beta}}^\lambda) \subseteq \mathcal{S}(\boldsymbol{\beta}^*)$ *with respect to* N*, when* ν *is set as* 0.005*.*

to 0.005, ℓ_1-PsLS fails to recover the governing PDE. On the other hand, (b) says that there exists a range of λ_N for which ℓ_1-PsLS can still recover a subset of $\boldsymbol{\beta}^*$, while the perfect signed-support recovery is difficult.

8. Discussion. We present future directions that can be further explored based on our ℓ_1-PsLS method.

1. Recall that our theory utilizes the equivalent kernel theory for Local-Polynomial regres-

sion [8], stating that the higher-order Local-Polynomial smoothing is asymptotically equivalent to higher-order kernel smoothing. Due to this construction, our theory cannot characterize the convergence behavior of signed-support recovery of ℓ_1-PsLS, when the number of observations is small. We conjecture that the uniform convergence rate of the Local-Polynomial estimator with exponential decay can be obtained in a non-asymptotic sense, by using a similar technique employed in [52]. They impose an assumption that the regression function belongs to the Hölder class. They manipulate the closed-form solution of the Local-Polynomial estimator so that the difference of the estimator and the regression function has a special form that can be controlled by the Bernstein's inequality. It would be an interesting research direction to see whether this technique can be employed in our setting.

2. The choice of the bandwidth parameter is essential in Local-Polynomial fitting, thereby having a significant impact on support recovery of PDE problem via ℓ_1-PsLS. It is worth noting that [10] employed the substitution method in [53] based on the asymptotic Mean Integrated Squared Error for the specific choices of the constant factors of the bandwidth parameter. However, the method is only limited to the local-quadratic estimator and is not applicable to our setting, which requires a higher-order smoothing estimator. In our numerical experiments, we choose the constant factors of bandwidth parameters h_N and w_M manually. It only provides an ad-hoc guidance of bandwidth selection. Developing a data-driven bandwidth selection procedure for ℓ_1-PsLS is a worthy topic for future research.

3. In practice, we need to set $P_{\max}$ large so as to avoid the model misspecification. Specifically, when $P_{\max}$ is set to be very large, the dimension of columns of $\widehat{\mathbf{F}}$ can be approximated as $K \approx \big(P_{\max}+1\big)^2$ in our problem setting. (Recall that we set $K = 1 + 2(P_{\max}+1) + \binom{P_{\max}+1}{2}$.) Under finite grid size NM, it is a possible scenario in which we have $K \gg NM$. Can we reduce the computational burdens in this case? As one possible direction, we can think of using the Sure Independence Screening (SIS) process [54] before solving ℓ_1-PsLS in (3.7). SIS is a dimension reduction technique before implementing variable selection algorithms, such as Lasso, SCAD, LARS, etc. In our case, for implementing SIS, we need to compute the marginal correlation between the response vector $\mathbf{u}_t$ and columns in $\widehat{\mathbf{F}}$, denoted as $\omega = \widehat{\mathbf{F}}^\top \widehat{\mathbf{u}}_t \in \mathbb{R}^K$. The paper [54] proved that with a certain choice of d, it is guaranteed that all the relevant predictors in $\widehat{\mathbf{F}}$ with $\widehat{\mathbf{u}}_t$ are included under regularity conditions on $\widehat{\mathbf{F}}$. Then, we may choose the largest d entries of the vector $|\omega|$, such that $K \gg NM \gg d$. The computational complexities of solving (3.7) via the well-known LARS algorithm [55] is known to be in the order of $\mathcal{O}\big(NMp \cdot \min(NM, p)\big)$, where p is set to be K before implementing the SIS and d after implementing the SIS. However, we need further studies on whether SIS will work well in the PDE identification problem, with theoretical guarantees. We leave this as a future work.

4. As one of the referees mentioned, the Theorem 4.1 cannot provide a guideline in prac-

tice, whether the selected model excludes crucial terms or even includes the irrelevant terms. To the best of the our knowledge, this is largely an open problem in the PDE identification context. From statisticians' viewpoint, we can suggest constructing a hypothesis testing, for $j \in \{1, 2, \ldots, K\}$, $H_0 : \widehat{\beta}_j^{\lambda} = 0$ v.s. $H_1 : \widehat{\beta}_j^{\lambda} \neq 0$. However, this is a challenging problem since we need to derive the distribution of the estimated coefficient $\widehat{\beta}_j^{\lambda}$ for each $j \in \{1, \ldots, K\}$. We are aware of the work [56] on constructing the confidence intervals of the Lasso estimator $\widehat{\beta}_j^{\lambda}$ under the classical *i.i.d.* centered normal error distribution. Nonetheless, this assumption is not applicable in our problem setting, and requires further investigations. We leave this problem as a future work.

5. It is worth noting that our paper is about model selection consistency of PDEs under noisy data and we consider the study on the estimation accuracy of the selected model is beyond the scope of our work. Nevertheless, it is still of importance to investigate whether the regression-based PDEs give a solution that closely resembles the original one. In practice, we suggest using the least-square estimate with the the selected features through ℓ_1-PsLS; that is, given that the ℓ_1-PsLS selects the true support set $\mathcal{S}$, then the least-square estimate has a form: $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}} := \left(\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{F}}_{\mathcal{S}}\right)^{-1}\widehat{\mathbf{F}}_{\mathcal{S}}^{\top}\widehat{\mathbf{u}}_t$. Note that $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ can avoid the bias introduced from λ_N and gives the consistent estimate than ℓ_1-PsLS. Although not reported in the paper, we verify that the least-square estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{LS}}$ works pretty well for the cases of KdV and viscous Burger's equations in section 7 in terms of estimation. We leave the study on the theoretical properties of this estimator as the future work. For more specific application with smaller data, there are related work with more refined model selection procedure, including [26] and [57]. We refer the readers these works and references therein.

Acknowledgements. The authors thank two anonymous reviewers, the Associate Editor, and the Chief-in-Editors for their thoughtful and constructive comments that greatly improved the presentation of this article.

REFERENCES

[1] Erwin Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6):1049, 1926.

[2] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *Journal of political economy*, 81(3):637–654, 1973.

[3] Jan Haskovec, Lisa Maria Kreusser, and Peter Markowich. ODE and PDE based modeling of biological transportation networks. *arXiv preprint arXiv:1805.08526*, 2018.

[4] Yves Achdou, Francisco J Buera, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Partial differential equation models in macroeconomics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 372(2028):20130397, 2014.

[5] Toshimitsu Musha and Hideyo Higuchi. Traffic current fluctuation and the Burgers equation. *Japanese journal of applied physics*, 17(5):811, 1978.

[6] 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.

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

[8] 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.

[9] Jianqing Fan. *Local polynomial modelling and its applications: monographs on statistics and applied probability 66*. Routledge, 2018.

[10] 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.

[11] 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.

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

[13] Markus Bär, Rainer Hegger, and Holger Kantz. Fitting partial differential equations to space-time dynamics. *Physical Review E*, 59(1):337, 1999.

[14] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[15] 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.

[16] Jinzhu Jia, Karl Rohe, and Bin Yu. The Lasso under poisson-like heteroscedasticity. *Statistica Sinica*, pages 99–118, 2013.

[17] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[18] Pradeep Ravikumar, Martin J Wainwright, and John D Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

[19] 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.

[20] Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

[21] 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.

[22] 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.

[23] 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.

[24] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.

[25] Alexandre Cortiella, Kwang-Chun Park, and Alireza Doostan. Sparse identification of nonlinear dynamical systems via reweighted ℓ_1-regularized least squares. *Computer Methods in Applied Mechanics and Engineering*, 376:113620, 2021.

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

[27] 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.

[28] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science Advances*, 3(4):e1602614, 2017.

[29] Hayden Schaeffer, Giang Tran, and Rachel Ward. Extracting sparse high-dimensional dynamics from limited data. *SIAM Journal on Applied Mathematics*, 78(6):3279–3295, 2018.

[30] Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

[31] David L Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE transactions on information theory*, 47(7):2845–2862, 2001.
[32] David L Donoho, Michael Elad, and Vladimir N Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on information theory*, 52(1):6–18, 2005.
[33] Arie Feuer and Arkadi Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.
[34] Emmanuel J Candes and Terence Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.
[35] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on information theory*, 52(2):489–509, 2006.
[36] Keith Knight and Wenjiang Fu. Asymptotics for Lasso-type estimators. *The Annals of statistics*, pages 1356–1378, 2000.
[37] Joel A Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE transactions on information theory*, 52(3):1030–1051, 2006.
[38] Peng Zhao and Bin Yu. On model selection consistency of Lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.
[39] Jean-Jacques Fuchs. Recovery of exact sparse representations in the presence of bounded noise. *IEEE Transactions on Information Theory*, 51(10):3601–3608, 2005.
[40] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *The Annals of statistics*, 34(3):1436–1462, 2006.
[41] Pradeep Ravikumar, Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Model selection in gaussian graphical models: High-dimensional consistency of ℓ_1-regularized mle. In *NIPS*, pages 1329–1336, 2008.
[42] Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101, 2010.
[43] 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.
[44] G Tusnády. A remark on the approximation of the sample df in the multidimensional case. *Periodica Mathematica Hungarica*, 8(1):53–55, 1977.
[45] Elias Masry. Multivariate local polynomial regression for time series: uniform strong consistency and rates. *Journal of Time Series Analysis*, 17(6):571–599, 1996.
[46] Yehua Li and Tailen Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *The Annals of Statistics*, 38(6):3321–3351, 2010.
[47] Bernard W Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *The Annals of Statistics*, pages 177–184, 1978.
[48] Mayur P Bonkile, Ashish Awasthi, C Lakshmi, Vijitha Mukundan, and VS Aswin. A systematic literature review of Burgers' equation with recent advances. *Pramana*, 90(6):69, 2018.
[49] OV Rudenko and SI Soluian. The theoretical principles of nonlinear acoustics. *MoIzN*, 1975.
[50] Katuro Sawada and Takeyasu Kotera. A method for finding n-soliton solutions of the KdV equation and KdV-like equation. *Progress of Theoretical Physics*, 51(5):1355–1367, 1974.
[51] Joseph Boussinesq. *Essai sur la théorie des eaux courantes*. Impr. nationale, 1877.
[52] Jean-Yves Audibert and Alexandre B Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of statistics*, 35(2):608–633, 2007.
[53] David Ruppert, Simon J Sheather, and Matthew P Wand. An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association*, 90(432):1257–1270, 1995.
[54] Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.
[55] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.
[56] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.
[57] Yuchen He, Sung Ha Kang, Wenjing Liao, Hao Liu, and Yingjie Liu. Robust PDE identification from noisy data. *arXiv preprint arXiv:2006.06557*, 2020.