IDENTIFIABILITY AND DATA-ADAPTIVE RKHS TIKHONOV REGULARIZATION IN NONPARAMETRIC LEARNING PROBLEMS

by

Qingci An

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March, 2023

Abstract

We provide an identifiability analysis for the learning problems (1) nonparametric learning of kernels in operators and (2) unsupervised learning of observation functions in state space models (SSMs). We show that in either case the function space of identifiability (FSOI) from the quadratic loss functional is the closure of a system-intrinsic data-adaptive reproducing kernel Hilbert space (SIDA-RKHS). We introduce a new method, the Data-Adaptive RKHS Tikhonov Regularization method (DARTR). The regularized estimator is robust to noise and converges as data refines. The effectiveness of DARTR is demonstrated through the following problems (1) nonparametric learning of kernels in linear/nonlinear/nonlocal operators and (2) homogenization of wave propagation in meta-material. We introduce a nonparametric generalized moment method to estimate non-invertible observation functions in nonlinear SSMs. Numerical results shows that the first two moments and temporal correlations, along with upper and lower bounds, can identify functions ranging from piecewise polynomials to smooth functions. The limitations, such as non-identifiability due to symmetry and stationary, are also discussed.

Primary Reader and Advisor: Fei Lu

Secondary Reader: Yanxun Xu

For my family.

Acknowledgements

I am deeply grateful to my advisor Fei Lu for his excellent supervision, and for guiding me into this beautiful subject. Fei has organized many learning seminars and introduced me to many conferences. These have tremendous influence on my mathematical taste. Fei suggested this thesis project to me, and has been unfailingly helpful since then. This thesis would have never existed without his guidance, encouragement and support.

I would like to thank Yannis Kevrekidis, Hans Lindblad, Xiong Wang, and Yanxun Xu for being in my thesis committee. The PhD thesis committee provided valuable feedbacks. I would like to thank their time and effort

I would like to thank Mauro Maggioni for the beneficial and enjoyable group meeting he organized. I would like to thank Jinchao Feng, Marie-José Kuffner, Christian Kümmerle, Patrick Martin, Zhong Ming, Sui Tang, Sichen Yang, and Felix Ye for their intriguing presentations. I have learned innumerable things from the sparkling conversations with them throughout graduate school.

I would like to thank Yue Yu for explaining to me her work [95, 96, 97]. Her insights on this subject have been invaluable to me.

I would like to thank all the fellow students and professors at the Department of Mathematics, Johns Hopkins University. I would like to thank Fei Lu, Emily Riehl and Yiannis Sakellaridis for organizing holiday dinners and parties. I would like to thank Quanjun Lang and Zehong Zhang for many inspirational discussions.

Special thanks to my college professors Cary Malkiewich and Charles Rezk, who encouraged me to pursue graduate studies. I would like to thank my college friend Yao Xiao for always supporting me. The time we spent on studying mathematics made my college life memorable.

Finally, I would like to thank my husband and my parents. This thesis would definitely be impossible without their continuous love and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In recent years, people have realized the potential of machine learning models in the scientific discovery of hidden physical laws of complex systems. In material science people seek to find material constitutive laws that best fit experimental data [63, 95] (Task A). Another example is the inference of an unknown observation function in a latent dynamical system from unlabeled data [27, 34, 45, 71] (Task B).

Task A can be formulated as learning a nonlocal operator that continuously maps the displacement field to the loading field. Beyond material science, the learning of operators between function spaces has broad applications in areas such as homogenization problems [60, 61, 97], fast PDE solvers [51, 58, 68, 69], and control problems [40, 57]. Motivated by these applications, an important inverse problem emerges: to learn the integral kernels in operators from data. Such kernel functions are resolution-invariant and reveal the law of nonlocal interaction. However, despite a long line of work on nonlocal models, there is limited theoretical characterization of the underlying inverse problems, even in the linear setting. The contribution of this thesis is filling

the gap by studying the linear inverse problem of nonparametric learning of kernels in operators.

Task B has been studied in many contexts including nonlinear system identification [9, 62], data assimilation [54], filtering and smoothing recursions [13], albeit typically only when observations are in the form of one, or a small number of, long trajectories, and in the case of invertible or smooth observations functions. State space models (SSMs) have been widely used to model such complex dynamics. When the observation function is invertible, its unsupervised regression is investigated by maximizing the likelihood for high-dimensional data [78]. However, in many applications, particularly those involving complex real-world physical processes, the observation functions are non-invertible or non-smooth. A case of particular interest in the present work is when the observation function is non-invertible or non-smooth. We introduce a new approach, the nonparametric generalized moment method, which does not require the invertibility and smoothness of the observation function. By this method, identifying a observation function from unlabeled data is equivalent to identifying a function-valued parameter in a operator that maps the generalized moments of the hidden process to these of the observation process. We will also emphasize the usefulness of many short trajectories (vs. few long trajectories), albeit both the theory and algorithms that we consider are generally applicable in a wide range of regimes. This study provides a first step in the unsupervised learning of latent dynamics from abundant unlabeled data.

Our goal is to infer such function-valued parameters in operators from data via nonparametric regression, when there is limited information to derive a parametric form. Three challenges are to be overcome. First, the function space of identifiability (FSOI) is yet to be specified properly, without which the inverse problem is ill-defined in the sense that there are multiple estimators fitting the data. Second, the estimator should converge in a proper function space as data resolution refines/data size increases, in order that it can be applied to problems and simulation tasks with

different grids or discretization methods and provides a guaranteed modeling accuracy. Third, the estimator should be robust to imperfect data.

To overcome these challenges, we introduce an exploration measure ρ, and provide an identifiability theory for the nonparametric learning of kernels in operators and the unsupervised learning of observation functions in SSMs. The exploration measure ρ quantifies the exploration of the kernel/observation function's variable by data. With this measure, we have an ambient function space $L^2(\rho)$ of learning. In either problems, the FSOI from the quadratic loss functional is the closure of a system-intrinsic data-adaptive reproducing kernel Hilbert space (SIDA-RKHS); it can be a proper subspace of $L^2(\rho)$. It is the first result on specifying a data-adaptive FSOI (see Lemma 3.2, 4.4 and Theorem 3.5, 4.3). It follows that the inverse problem is ill-defined beyond the FSOI and is ill-posed in the FSOI (see details in Chapter 3.1 and 4.3). To overcome the ill-posedness and to ensure the learning take place inside the FSOI, we introduce a novel data-adaptive RKHS Tikhonov regularization (DARTR) method that uses the norm of the SIDA-RKHS (see Chapter 3.2). Finally, We validate the theory and the proposed algorithm on a number of benchmark problems, including various synthetic datasets and a real-world dataset. Numerical results in Chapter 3.3 show that the proposed algorithm in Chapter 3.2 provides stable and convergent estimations of kernel in linear/nonlinear/nonlocal operators. The rates of convergence are robust to different levels of white noise in data, while the common Tikhonov/ridge regularizers with l^2 or L^2-norm fail this task. It also successfully learns a homogenized model for the wave propagation in a heterogeneous solid, revealing the unknown resolution-invariant governing laws from real-world data at microscale (see Chapter 3.4). The identifiability analysis and the DARTR method are applicable to general linear inverse problems that minimize quadratic loss functionals. It is particularly useful when the data depends non-locally on the unknown function.

1. Nonparametric learning of function-valued parameters

We consider the linear inverse problem of identifying function-valued parameters in operators in the form of

$$R_\phi : \mathbb{X} \to \mathbb{Y},\ R_\phi[u] = v, \tag{1.1}$$

from data pairs

$$\mathcal{D} = \{(u_k, v_k)\}_{k=1}^N, \quad (u_k, v_k) \in \mathbb{X} \times \mathbb{Y}, \tag{1.2}$$

where $\mathbb{X}$ and $\mathbb{Y}$ are Hilbert spaces. Our goal is to find a function-valued parameter ϕ in such operator R_ϕ that best fits the data pairs $\{(u_k, v_k)\}_{k=1}^N$. The operator R_ϕ can be either linear or nonlinear in u. We assume that it depends linearly on ϕ.

We aim for an estimation of the function-valued parameter ϕ that (1) converges in a proper function space to the true function as data mesh refines/data size increases in the synthetic setting and (2) robust enough to treat imperfect data in order to be applicable in real applications.

In this thesis, we focus on the following two problems that can be reduced to the above general learning problem.

Nonparametric learning of kernels in operators. In Chapter 2 – 3, we focus on such operators in the form of

$$R_\phi[u](x) = \int_\Omega \phi(|y|) g[u](x, y) dy,\ \forall x \in \Omega, \tag{1.3}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded connected open set, ϕ is a radial kernel, $\mathbb{X} = H_0^1(\Omega)$ and $\mathbb{Y} = L^2(\Omega)$. The functional g, which may depend on the derivatives of u, is assumed known and it specifies the form of the operator.

Unsupervised learning of observation functions in SSMs. In Chapter 4, we investigate the unsupervised learning of non-invertible observation functions in nonlinear SSMs. With the proposed nonparametric generalized moment method method in Chapter 4.2, identifying the observation function f_* is equivalent to identifying a function-valued parameter ϕ in a operator R_ϕ that maps the generalized moments of the hidden process $f(X_t)$ to these of the observation process Y_t. That is, we consider an operator in the form of $R_\phi[u] = \xi(\phi(u))$, and we want to identify the function-valued parameter ϕ from the known information $u_t = X_t$ and the empirical approximations of the moments $v_t = \xi(Y_t)$, $t \in \{t_0, t_1, \cdots, t_N\}$. For the sake of computation efficiency, the generalized moment functional ξ is often chosen as low-order moments. Matching the first moments yields to an operator R_ϕ linear in ϕ

$$R_\phi[u] = \mathbb{E}\left[\phi(u)\right], \tag{1.4}$$

We will discuss matching second moments and temporal correlations in Chapter 1.4 and more technical details will be presented in Chapter 4.2. In this and the next section, we only consider matching the first moments in order to provide our main results on identifiability analysis from the quadratic loss functionals.

Quadratic loss functionals. In either problems, we construct a variational estimator that minimizes the mean square error,

$$\widehat{\phi} = \underset{\phi \in \mathcal{H}}{\arg\min}\, \mathcal{E}(\phi), \qquad \text{where } \mathcal{E}(\phi) = \frac{1}{N}\sum_{k=1}^{N} \|R_\phi[u_k] - v_k\|_{\mathbb{Y}}^2. \tag{1.5}$$

The hypothesis space $\mathcal{H}$ is to be selected adaptive to data. Note that the loss functional $\mathcal{E}(\phi)$

is quadratic in ϕ since the operator R_ϕ depends linearly on ϕ. Thus, the minimizer of the loss functional is the least square estimator. Suppose the hypothesis space $\mathcal{H}_n$ is the linear span of the basis functions $\{\phi_i\}_{i=1}^n$. Let $\phi \in \mathcal{H}_n$ and write $\phi = \sum_{i=1}^n c_i \phi_i$. Noticing that $R_\phi = \sum_{i=1}^n c_i R_{\phi_i}$, we can write the loss functional in (1.5) as

$$\mathcal{E}(c) = \mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^v, \tag{1.6}$$

where $C_N^v = \frac{1}{N}\sum_{k=1}^N \langle v_k, v_k\rangle_{\mathbb{Y}}$ and the normal matrix $\overline{A}_n$ and vector $\overline{b}_n$ are given by

$$\overline{A}_n(i,j) = \frac{1}{N}\sum_{k=1}^N \langle R_{\phi_i}[u_k], R_{\phi_j}[u_k]\rangle_{\mathbb{Y}}, \quad \overline{b}_n(i) = \frac{1}{N}\sum_{k=1}^N \langle R_{\phi_i}[u_k], v_k\rangle_{\mathbb{Y}}. \tag{1.7}$$

The least square estimator is computed directly from the minimizer of the quadratic function $\mathcal{E}(c)$:

$$\widehat{\phi}_{\mathcal{H}_n} = \sum_{i=1}^n \widehat{c}_i \phi_i \quad \text{and} \qquad \widehat{c} = \overline{A}_n^{-1}\overline{b}_n, \tag{1.8}$$

where $\overline{A}_n^{-1}$ is the inverse of $\overline{A}_n$ or Moore–Penrose pseudo-inverse when $\overline{A}_n$ is singular.

In nonparametric learning, it is often necessary to select a relatively large hypothesis space to make the model flexible enough. However, the large hypothesis space often leads to a severely ill-conditioned normal matrix $\overline{A}_n$. Without proper regularization, the estimator often oscillates largely from data to data due to overfitting. Thus, identifiability analysis and regularization is crucial for obtaining noise-robust and convergent estimators.

2. Identifiability

The main theme in the identifiability theory is to find the largest function space in which the loss functional has a unique minimizer. This is a generalization of the uniqueness of minimizer of a loss function in parametric inference (see [10, page 431] and [25]).

Definition 1.1 (Function space of identifiability)**.** *The function space of identifiability (FSOI) is the largest subspace of $L^2(\rho)$ in which the true function ϕ_{true} is the unique minimizer of the loss functional $\mathcal{E}$ with continuous noiseless data.*

Finding the FSOI of the nonparametric learning of function-valued parameters in operators is fundamentally different from that of the classical nonparametric regression. Recall that the classical nonparametric regression learns a function $Y = \phi(X)$ from random samples $\{(X_i, Y_i)\}$ from the joint distribution of (X, Y). The FSOI in classical regression is $L^2(\rho)$ with ρ being the distribution of X, and the optimal estimator is the conditional expectation. In nonparametric learning of function-valued parameters in operators, the loss functional has a unique minimizer in a Hilbert space if and only if its Fŕechet derivative is invertible in the Hilbert space; thus, the main task is to find such function space [52, 56, 65].

We show that the function space of identifiability derived by the loss functional in (1.5) is the $L^2(\rho)$-closure of a system-intrinsic data-adaptive reproducing kernel Hilbert space. This space is the image of the square root of the Fréchet derivative of the loss functional, which is a compact operator. Therefore the inverse problem is ill-posed since it requires the inversion of a compact operator.

More specifically, we show that the quadratic error functional of the form (1.5) in either prob-

lems can be rewrite as

$$\mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}}\phi, \phi\rangle_{\mathbb{Y}} - 2\langle \phi_N^v, \phi\rangle_{\mathbb{Y}} + C_N^v,$$

where $\mathcal{L}_{\overline{G}}$ is a system-intrinsic data-adaptive compact and positive semi-definite integral operator and $\phi_N^v \in L^2(\rho)$ is the Riesz representation of the bounded linear functional

$$\langle \phi_N^v, \psi\rangle_{L^2(\rho)} = \frac{1}{N}\sum_{k=1}^{N}\langle R_\psi[u_k], v_k\rangle_{\mathbb{Y}},\ \forall \psi \in L^2(\rho). \tag{1.9}$$

The Fréchet derivative of $\mathcal{E}(\phi)$ in $L^2(\rho)$ is $\nabla\mathcal{E}(\phi) = 2(\mathcal{L}_{\overline{G}}\phi - \phi_N^v)$. The function space of identifiability is $H = \overline{\text{span}\{\psi_i\}}$ with closure in $L^2(\rho)$, where $\{\psi_i\}$ are eigenfunctions of the integral operator $\mathcal{L}_{\overline{G}}$ with positive eigenvalues. Furthermore, the unique minimizer of $\mathcal{E}(\phi)$ in H is $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^v$ if $\phi_N^v \in \mathcal{L}_{\overline{G}}(L^2(\rho))$. When zero is an eigenvalue of $\mathcal{L}_{\overline{G}}$, this function space is a proper subspace of $L^2(\rho)$ and the loss functional has multiple minimizers in $L^2(\rho)$. Thus, the inverse problem is well-defined only on H. In particular, if the data is perfect and generated from a true function ϕ_{true}, we have $\phi_N^v = \mathcal{L}_{\overline{G}}\phi_{true}$ and $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^v = \phi_{true}$. The Fréchet derivative of $\mathcal{E}$ in the SIDA-RKHS $H_G = {\mathcal{L}_{\overline{G}}}^{1/2}(L^2(\rho))$ with $\overline{G}$ as the reproducing kernel is $\nabla^{H_G}\mathcal{E}(\phi) = 2({\mathcal{L}_{\overline{G}}}^2\phi - \mathcal{L}_{\overline{G}}\phi_N^v)$. Its zero leads to another estimator $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-2}\mathcal{L}_{\overline{G}}\phi_N^v$ if $\phi_N^v \in \mathcal{L}_{\overline{G}}(L^2(\rho))$.

In the unsupervised learning of observation function in SSMs, the proposed nonparametric generalized moment method works well when the densities of the hidden process X_t vary appreciably in time to yield a large FSOI, whose distance to the true observation function is small. In this case, our algorithm 2 leads to a noise-robust and convergent estimator as the sample size increases. When the hypothesis space is larger than the FSOI, the quadratic loss functional may have multiple minimizers. The constraints of upper and lower bounds, as well as the quartic loss functionals from matching the second moments and temporal correlations, can help identifying the

observation function. However, identifiability may still not hold due to symmetry and/or stationarity (see Section 3.2, Chapter 4). We analytically compute the reproducing kernels associated with the SIDA-RKHSs when the hidden process X_t is the Brownian motion or the stationary Ornstein-Uhlenbeck (OU) process (see Example 4.10 and Example 4.11 in Chapter 4.3).

3. Regularization

Regularization plays a crucial role in inverse and machine learning problems that aim to construct robust generalizable model. Various regularization methods have been introduced to prevent overfitting in such ill-posed inverse problems. The idea is to add a penalty term to the loss functional:

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \mathcal{R}(\phi), \tag{1.10}$$

where $\mathcal{R}(\phi)$ is a regularization term and λ is a hyperparameter which controls the impact of the regularization. By adding a penalty term to the loss functional, regularization controls the complexity of the model and prevents it from fitting noise in the data. Various penalty terms have been proposed, including the Euclidean norm $\mathcal{R}(\phi) = \|c\|^2$ for $\phi = \sum_{i=1}^n c_i \phi_i$ in classical Tikhonov regularization [37, 88], the RKHS norm $\mathcal{R}(\phi) = \|\phi\|_H^2$ with H being a reproducing kernel Hilbert space with an artificial reproducing kernel, the total variation norm $\mathcal{R}(\phi) = \|\phi'\|_{L^1}$ in Rudin–Osher–Fatemi method or the L^1 norm $\mathcal{R}(\phi) = \|\phi\|_{L^1}$ in LASSO.

Whereas each of these penalty terms has their specific reasoning and applications, none of them take into account of the function space of identifiability by the loss functional (1.5), only in which the inverse problem is well-defined. The major novelty of DARTR method is the construction of a SIDA-RKHS, whose reproducing kernel is encoded in the loss functional. DARTR takes the norm

of this RKHS as the the penalty norm of regularization, and ensures the regularized estimator lies in the function space of identifiability.

Numerical results in Chapter 3.3 and 3.4 show that the DARTR method yields a noisy-robust convergent estimator of the kernel as data resolution refines. The estimators regularized by SIDA-RKHS norm are more accurate than those of regularized by the l^2 or L^2 norm when the regularization hyperparameter λ is properly selected. The SIDA-RKHS regularizer robustly leads to estimators converging at a consistent rate for all levels of noises for linear/nonlinear/nolocal operator, while the other two regularizers cannot. In comparison to regularization by l^2 or L^2 norm, the additional computational cost is negligible.

Our regularization method is inspired by the kernel flow method that learns hyper-parameters of the reproducing kernel [16, 35, 73], but our reproducing kernel is adaptive to the system and the data.

4. Nonparametric generalized moment method

We consider the following state space model for processes (X_t, Y_t) in $\mathbb{R} \times \mathbb{R}$:

$$\text{State equation:} \qquad dX_t = a(X_t)dt + b(X_t)dB_t, \qquad \text{with } a, b \text{ are known;} \tag{1.11}$$

$$\text{Observation equation:} \qquad Y_t = f_*(X_t), \qquad \text{with } f_* \text{ unknown.} \tag{1.12}$$

Here B_t is the standard Brownian motion, the drift function $a(x)$ and the diffusion coefficient $b(x)$ are given, satisfying the linear growth and global Lipschitz conditions. We assume that the initial distribution of X_{t_0} is given. Thus, the distribution of the state process (X_t) is known.

Our goal is to estimate the unknown observation function f_* from data consisting of a large

ensemble of trajectories of the process Y_t, denoted by $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$, where m indexes trajectories, and $t_0 < \cdots < t_L$ are the times at which the observations are made. In particular, there are no pairs (X_t, Y_t) being observed, so in the language of machine learning this may be considered an unsupervised learning problem.

We estimate the observation function f_* by matching generalized moments, while constraining the estimator to a suitably chosen finite-dimensional hypothesis space, whose dimension depends on the number of observations. The estimator $\widehat{f}$ minimizes the discrepancy between the the first two moments and one-step temporal correlations of functionals of the process $\widehat{f}(X_t)$ and the empirical ones (computed from data) of $Y_t = f_*(X_t)$, with upper and lower pointwise constraints estimated from data. Matching the first moments yields a quadratic loss functional, and matching the second moments and temporal correlations yields two quartic loss functionals. We summarize the proposed method in Algorithm 2. The algorithm is suitable for large sets of unlabeled data. Moreover, it can deal with challenging cases when the observation function is nonlinear and non-invertible. Numerical examples show the proposed algorithm can identify smooth/non-smooth functions and tolerate considerable observation noise (see Chapter 4.4). The estimation error caused by the noise is at the scale of the sampling error, which is negligible when the sample size is large.

The method we propose has several significant strengths: (1) the generalized moments do not require the invertibility of the observation function; (2) low-order generalized moments tend to be robust to additive observation noise; (3) generalize moments avoid the need of local constructions, since they depend on the entire distribution of the latent and observed processes; (4) our nonparametric approach does not require *a priori* information about the observation function, and it can deal with both regular and piecewise regular functions; (5) the method is computationally efficient because the moments need to be estimated only once, and the computation is can performed in

parallel.

We note that the method we propose readily extends to multivariate state space models, with the main statistical and computational bottlenecks coming from the curse of dimensionality in the representation and estimation of a higher-dimensional observation function in terms of basis functions.

5. Related work

Latent dynamical systems. The estimation of the unknown observation function and of the latent dynamics from unlabeled data has been considered in [27, 34, 45, 71] and references therein. Most classical approaches focus on estimating the parameters in the space-state model from a single trajectory of the observation process, by expectation-maximization methods maximizing the likelihood, or Bayesian approaches [9, 13, 28, 46, 62], with the recent studies estimating the coefficients in a kernel representation [89] or the coefficients of a pre-specified set of basis functions [87]. The recent work [94] estimates a slow manifold (and effective equations on it), image under a nonlinear but invertible map of a latent space where slow and fast variables in a slow-fast system of SDEs are independent and orthogonal, using short bursts of trajectories; see discussions and references therein for motivations, applications and related works. Our framework combines nonparametric learning [20, 33] with the generalized moment method, that is mainly studied in the setting of parametric inference [76, 77, 85].

Nonlocal operators. Models with nonlocal operators have received increasing attention, since they can describe physical phenomena involving nonlocal or long-range dependence that classical differential operators fail to capture and provide a powerful model for a large class of complex engineering and scientific applications including nonlocal and fractional diffusion [2, 3, 12, 17, 22,

23, 84, 90, 93], homogenization problems [60, 97, 61, 95], fast PDE solvers [51, 57, 68, 69], control problems [40, 69], subsurface transport [7, 47, 48, 80, 81], multi-agent systems with nonlocal interaction [52, 66, 67], phase transitions [5, 15, 21], nonlocal network in machine learning [92, 59] and image processing [11, 29, 30, 38, 49, 64]. The inverse problem for nonlocal diffusion has been studied in [43, 55] from a single solution. To discover nonlocal physical laws from data, a parametric learning approach has been proposed in [95, 96], where the coefficients of Bernstein polynomials are learnt with physics-based constraints and a Tikhonov regularization. Beyond the linear nonlocal model and the regression methods, nonlocal operators were further combined with neural networks, and nonlocal kernel networks were developed for learning maps between high-dimensional variables in dynamical systems or function spaces [57, 58, 98]. An attractive feature of these nonlocal kernel/operator learning methods is the generalizability among approximations corresponding to different underlying levels of resolution and discretization. However, as seen in [58, 95, 97, 98], none of them yield estimator convergence when trained on finer resolution, and the test error may even increase, due to the ill-posedness of the inverse problem. In this work, we tackle this issue by introducing a new regularization method based on a system-intrinsic data-adaptive RKHS in a nonparametric learning approach.

Data-dependent function spaces. Data-dependent strategies have been explored in the context of classical nonparametric regression, such as data-dependent hypothesis space with an l^1 regularizer [82, 91] and data-dependent early stopping rule [79]. While all strategies achieve data-dependent regularization, our DARTR method is tailored for the learning of kernels in operators, for which the function space of identifiably places a fundamental role.

Bayesian perspective and Zellner's g-prior. In a Bayesian perspective, the Tikhonov regularization can be interpreted as a Gaussian prior distribution with a covariance matrix corresponding

to the penalty term. In this perspective, our SIDA-RKHS norm coincides with the Zellner's g-prior [1, 6, 99] when the data has additive white noise.

Relation to classical regression. When the data $\{(u_k, v_k)\}_{k=1}^N$ are scalars instead of functions and the operator $R_\phi(u) = \phi(u)$, we get back to the classical regression problem [19, 33]. Our data-adaptive RKHS reduces to the empirical $L^2(\rho)$ space with ρ being the distribution of data $\{u_k\}$, and the DARTR regulation reduces to the classical L^2 Tikhonov/ridge regularization.

Relation to functional data analysis. The problem of learning kernel in operators can be viewed as a problem in functional data analysis (see [39, 44]), where the data are samples from distributions on function spaces. Our DARTR method is applicable to this setting. However, this study focuses on the situation of limited deterministic data (with only a few pairs of data) and on discovering an intrinsic low-dimensional kernel function.

Chapter 2

The inverse problem of learning kernels in operators

1. Introduction and related work

The learning of kernel functions in operators is such a problem: given data $\{(u_k, v_k)\}_{k=1}^N$ in suitable function spaces, we would like to learn the kernel function ϕ in the operator $R_\phi(u) = f$ to best fit the data. Such a need for learning operators between function spaces has become vital in applications ranging from integral operators solving PDEs [31, 51, 57, 58], nonlinear operators in mean-field equation of interacting particle systems [53, 66], homogenized nonlocal operators [61, 95, 97], just to name a few. Since there is often limited information to derive a parametric form, the kernel has to be learnt in a nonparametric fashion. More importantly, the goal is a consistent estimator that converges as data mesh refines and is robust to noise in data. Without proper regularization, the estimator often oscillates largely from data to data due to overfitting. Thus, regularization is crucial for the discovery of the best kernel.

We present a data-adaptive RKHS Tikhonov regularization (DARTR) method for the linear inverse problem of learning of kernels in operators from data. That is, the operator $R_\phi(u)$, which can be either linear or nonlinear in u, depends linearly on the kernel ϕ. We learn the kernel by nonparametric regression that minimizes a loss functional of the mean square error. With DARTR, our nonparametric regression algorithm produces an estimator that converges as the data mesh refines and the rate of convergence is robust to different levels of white noise in data.

The major novelty of this method is the construction of a system (operator) intrinsic data-adaptive (SIDA) RKHS, whose reproducing kernel is encoded in the loss functional. Additionally, we introduce a novel exploration measure quantifying the exploration of the kernel by the data, and it allows a unified framework to treat SIDA-RKHS with either discrete or continuous functions. DARTR takes the norm of this RKHS as the the penalty norm of regularization, and ensures the learning to take place in the function space of identifiability.

2. The inverse problem and the need of regularization

In this study, we focus on such operators in the form of

$$R_\phi[u](x) = \int_\Omega \phi(|y|)g[u](x,y)dy, \ \forall x \in \Omega. \tag{2.1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded connected open set, ϕ is a radial function-valued parameter. Given data

$$\mathcal{D} = \{(u_k, v_k)\}_{k=1}^N, \quad (u_k, v_k) \in H_0^1(\Omega) \times L^2(\Omega), \tag{2.2}$$

our goal is to find a function-valued parameter ϕ so that R_ϕ best fits the data pairs $\{(u_k, v_k)\}_{k=1}^N$ in the form (2.1).

The functional g, which may depend on the derivatives of u, is assumed to be known and it specifies the form of the operator. Examples are as follows: (see more details in Chapter 3.3)

- R_ϕ is an *integral operator* with $g[u](x,y) = u(x+y)$ and ϕ is called an integral kernel.
- R_ϕ is a *nonlinear operator* with $g[u](x,y) = u'(x+y)u(x)$ and ϕ is called an interaction kernel in mean-field equation of interacting particles.
- R_ϕ is a *nonlocal operator* with $g[u](x,y) = u(x+y) - u(x)$ with ϕ called a nonlocal kernel.

These inverse problems share three common features: First, the pointwise values of the function ϕ are undetermined from data, because the data depends on ϕ non-locally. Also, the support of ϕ is unknown and is to be learnt from data. Second, the data are discrete and can be noisy. Thus, the inverse problem has to overcome the numerical error in the approximation of integrals, as well as the measurement error. Third, the inverse problem extends to a homogenization problem where the operator aims to fit the data that are not generated from the equation (2.1). In this case, the inverse problem has to overcome the model error to identify a best fit.

For simplicity of representation, in the next two sections on learning theory, we consider the loss functional mainly for continuous data. All the arguments apply directly to discrete data by replacing the integrals with Riemann sum or another numerical integrator.

We construct a variational estimator that minimizes the mean square error:

$$\widehat{\phi} = \arg\min_{\phi \in \mathcal{H}} \mathcal{E}(\phi), \qquad \text{where } \mathcal{E}(\phi) = \frac{1}{N} \sum_{k=1}^{N} \int_\Omega |R_\phi[u_k](x) - v_k(x)|^2 dx, \tag{2.3}$$

where the hypothesis space $\mathcal{H}$ is to be selected adaptive to data. Suppose the hypothesis space $\mathcal{H}$ is $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ with basis functions $\{\phi_i\}_{i=1}^n$. Then for each $\phi = \sum_{i=1}^n c_i \phi_i \in \mathcal{H}_n$, noticing

that $R_\phi = \sum_{i=1}^n c_i R_{\phi_i}$, we can write the loss functional in (2.3) as

$$\mathcal{E}(c) = \mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^v, \tag{2.4}$$

where $C_N^v = \frac{1}{N}\sum_{k=1}^N \int_\Omega |v_k(x)|^2 dx$ and the normal matrix $\overline{A}_n$ and vector $\overline{b}_n$ are given by

$$\overline{A}_n(i,j) = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle, \quad \overline{b}_n(i) = \frac{1}{N}\sum_{k=1}^N \int_\Omega R_{\phi_i}[u_k](x) v_k(x) dx, \tag{2.5}$$

where $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ is the bilinear form defined by

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \frac{1}{N}\sum_{k=1}^N \int_\Omega R_\phi[u_k](x) R_\psi[u_k](x) dx. \tag{2.6}$$

The least square estimator is computed directly from the minimizer of the quadratic function $\mathcal{E}(c)$:

$$\widehat{\phi}_{\mathcal{H}_n} = \sum_{i=1}^n \widehat{c}_i \phi_i \quad \text{and} \qquad \widehat{c} = \overline{A}_n^{-1} \overline{b}_n, \tag{2.7}$$

where $\overline{A}_n^{-1}$ is the inverse of $\overline{A}_n$ or Moore–Penrose pseudo-inverse when $\overline{A}_n$ is singular.

A major challenge is to find an optimal estimator capable of avoiding either under-fitting or over-fitting, being robust with imperfect data and model error and in particular, converging in synthetic tests when the data mesh refines. Unfortunately, this is an ill-posed inverse problem (see Chapter 3.1) and the normal matrix $\overline{A}_n$ is often highly ill-conditioned or singular. As a result, the estimator in (2.7) oscillates largely and fails to converge when the data mesh refines. In next chapter, we introduce a regularization method, DARTR, based on the identifiability analysis. The regularized estimator is accurate and robust to both numerical error due to discrete data and

noise in data; the regularized estimator converges at a consistent rate as the data mesh refines, outperforming two baseline regularizers using l^2 and L^2 norms, even after they take into account the function space of identifiability by an additional projection.

Chapter 3

Identifiability and Data-Adaptive RKHS Tikhonov regularization (DARTR)

1. Identifiability analysis

We aim to find the function space on which the quadratic loss functional has a unique minimizer. In other words, we seek the function space in which the Fréchet derivative of the loss functional is invertible. Using the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in (2.6), we can rewrite the loss functional in (2.3) as

$$\mathcal{E}(\phi) = \frac{1}{N}\sum_{k=1}^{N}\int R_\phi[u_k](x)^2 dx - 2\frac{1}{N}\sum_{k=1}^{N}\int R_\phi[u_k](x)v_k(x)dx + C_f, \tag{3.1}$$

where $C_N^v = \frac{1}{N}\sum_{k=1}^{N}\int |v_k(x)|^2 dx$. However, there is no function space for ϕ yet. To start with, we introduce two key elements: a data-adaptive exploration measure that leads to a default function space of learning and an integral operator which plays a crucial role in our DARTR. Throughout this section, we assume continuous data to simplify the notation. All the integrals will be

numerically approximated from discrete data in the next section.

Assumption 3.1. *The data* $\mathcal{D} = \{u_k, v_k\}_{k=1}^N$ *in* (2.2) *are continuous with compact support.*

The exploration measure. We introduce first a measure that quantifies the exploration of the independent variable of ϕ by the data. Given data in (2.2), we define an empirical measure

$$\rho(dr) = \frac{1}{ZN} \sum_{k=1}^{N} \int_\Omega \int_\Omega \delta(|y| - r) \, |g[u_k](x, y)| \, dxdy, \tag{3.2}$$

where $Z = \int_0^\infty \frac{1}{N} \sum_{k=1}^N \int_\Omega \int_\Omega \delta(|y| - r) \, |g[u_k](x, y)| \, dxdydr$ is the normalizing constant. By definition, this measure reflects the weight being put by the loss function on $|y|$ through the data $\{g[u_k](x, y)\}_{k=1}^N$.

The exploration measure plays an important role in the learning of the function ϕ. Its support is the region inside of which the learning process ought to work and outside of which we have limit information from the data to learn the function ϕ. Thus, it defines a default function space of learning: $L^2(\rho)$.

An integral operator. The loss functional's Fréchet derivative in $L^2(\rho)$ comes directly from the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in (2.6). To see this, we rewrite the bilinear form as

$$\begin{aligned} \langle\!\langle \phi, \psi \rangle\!\rangle =& \frac{1}{N} \sum_{k=1}^{N} \int \left[\int \int \phi(|z|)\psi(|y|) g[u_k](x, z) g[u_k](x, y) dydz \right] dx \\ =& \int_0^\infty \int_0^\infty \phi(r)\psi(s) G(r, s) drds = \int_0^\infty \int_0^\infty \phi(r)\psi(s) \overline{G}(r, s) \rho(dr)\rho(ds), \end{aligned} \tag{3.3}$$

where the second-to-last equation follows from a change of order of integration and a change of

variables to polar coordinates with the integral kernel G given by

$$G(r,s) = \frac{1}{N}\sum_{k=1}^{N}\int_{|\eta|=1}\int_{|\xi|=1}\left[\int g[u_k](x,r\xi)g[u_k](x,s\eta)dx\right]d\xi d\eta, \tag{3.4}$$

for $r, s \in \text{supp}(\rho)$ and $G(r,s) = 0$ otherwise. Here the last equality is a re-weighting by ρ with

$$\overline{G}(r,s) = \frac{G(r,s)}{\rho(r)\rho(s)}, \tag{3.5}$$

where, by an abuse of notation, we also use $\rho(r)$ to denote the density of the probability measure ρ defined in (3.2).

The next lemma shows that $\overline{G}$ defines a positive semi-definite integral operator.

Lemma 3.2 (The integral operator)**.** *Under Assumption* 3.1*, the integral kernel $\overline{G}$ is positive semi-definite and the integral operator $\mathcal{L}_{\overline{G}} : L^2(\rho) \to L^2(\rho)$*

$$\mathcal{L}_{\overline{G}}\phi(r) = \int_0^\infty \phi(s)\overline{G}(r,s)\rho(s)ds \tag{3.6}$$

is compact and positive semi-definite. Further more, for any $\phi, \psi \in L^2(\rho)$,

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \langle \mathcal{L}_{\overline{G}}\phi, \psi\rangle_{L^2(\rho)}; \tag{3.7}$$

Proof of Lemma 3.2. Recall that a bi-variate function $\overline{G}(r,s)$ is positive semi-definite if the sum $\sum_{i=1}^m \sum_{j=1}^m c_i c_j \overline{G}(r_i, r_j) \geq 0$ for any $(c_1, \ldots, c_m) \in \mathbb{R}^m$ and any $\{r_j\}_{j=1}^m \subset \mathbb{R}^d$ (see Appendix A.1).

Using (3.4) and (3.5), we have

$$\begin{aligned}\sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \overline{G}(r_i, r_j) =& \frac{1}{N}\sum_{k=1}^{N}\int_{|\eta|=1}\int_{|\xi|=1}\left[\int \sum_{i=1}^{m}\sum_{j=1}^{m} c_i c_j \frac{g[u_k](x, r_i\xi) g[u_k](x, r_j\eta)}{\rho(r_i)\rho(r_j)} dx\right] d\xi d\eta \\ =& \frac{1}{N}\sum_{k=1}^{N}\int_{|\eta|=1}\int_{|\xi|=1}\left[\int \left|\sum_{i=1}^{m} c_i \frac{g[u_k](x, r_i\xi)}{\rho(r_i)}\right|^2 dx\right] d\xi d\eta \geq 0.\end{aligned}$$

Thus $\overline{G}$ is positive semi-definite. The operator $\mathcal{L}_{\overline{G}}$ is compact because $\overline{G} \in L^2(\rho \times \rho)$, which follows from the fact that each u_k is bounded (thus, $\overline{G}$ is also bounded). Also, since $\overline{G}$ is positive semi-definite, so is $\mathcal{L}_{\overline{G}}$. The equation (3.7) follows from (3.3). □

The next lemma provides an operator characterization of the RKHS with $\overline{G}$ as the reproducing kernel [4]. This RKHS is system(the operator R_ϕ) intrinsic data adaptive (SIDA), and we refer it as SIDA-RKHS. It is the data adaptive RKHS in our DARTR.

Lemma 3.3 (The SIDA-RKHS)**.** *Assume Assumption* 3.1*. Then the following statements hold.*

(a) The RKHS H_G with $\overline{G}$ as the reproducing kernel satisfies $H_G = \mathcal{L}_{\overline{G}}^{1/2}(L^2(\rho))$ and its inner product satisfies $\langle \phi, \psi \rangle_{H_G} = \langle \mathcal{L}_{\overline{G}}^{-1/2}\phi, \mathcal{L}_{\overline{G}}^{-1/2}\psi \rangle_{L^2(\rho)}$ for any $\phi, \psi \in H_G$.

(b) The eigen-functions of $\mathcal{L}_{\overline{G}}$, denoted by $\{\psi_i, \psi_j^0\}_{i,j}$ with $\{\psi_i\}$ corresponding to positive eigenvalues $\{\lambda_i\}$ in decreasing order and $\{\psi_j^0\}$ corresponding to zero eigenvalues (if any), form an orthonormal basis of $L^2(\rho)$ and λ_i converges to 0. Furthermore, for any $\phi = \sum_i c_i\psi_i$, we have

$$\langle\!\langle \phi, \phi \rangle\!\rangle = \sum_i \lambda_i c_i^2, \quad \|\phi\|_{L^2(\rho)}^2 = \sum_i c_i^2, \quad \|\phi\|_{H_G}^2 = \sum_i \lambda_i^{-1} c_i^2, \tag{3.8}$$

where the last equation is restricted to $\phi \in H_G$.

(c) For any $\phi \in L^2(\rho)$ and $\psi \in H_G$, we have

$$\langle \phi, \psi \rangle_{L^2(\rho)} = \langle \mathcal{L}_{\overline{G}} \phi, \psi \rangle_{H_G}, \quad \langle\!\langle \phi, \psi \rangle\!\rangle = \langle \mathcal{L}_{\overline{G}}{}^2 \phi, \psi \rangle_{H_G}. \tag{3.9}$$

Proof of Lemma 3.3. Part (a) is a standard operator characterization of the RKHS H_G (see Appendix A.1).

For Part (b), since the operator $\mathcal{L}_{\overline{G}}$ is symmetric positive semi-definite and compact as shown in Lemma 3.2, the eigenfunctions are orthonormal and the eigenvalues decay to zero. The first equation in (3.8) follows from (3.7) and the second equation follows from the orthonormality of the eigenfunctions. At last, if $\phi \in H_G$, by the characterization of the inner product of H_G in Part (a), we have the third equation in (3.8).

The first equality in Part (c) follows from Part (a) and that $\mathcal{L}_{\overline{G}}{}^{-1/2}$ is self-adjoint, which imply that $\langle \mathcal{L}_{\overline{G}} \phi, \psi \rangle_{H_G} = \langle \mathcal{L}_{\overline{G}}{}^{1/2} \phi, \mathcal{L}_{\overline{G}}{}^{-1/2} \psi \rangle_{L^2(\rho)} = \langle \phi, \psi \rangle_{L^2(\rho)}$. The second equality in (3.9) follows from the first equality and (3.7). □

Remark 3.4. *The space $L^2(\rho)$ can be a discrete vector space with the function ϕ defined only on finitely many points $\{r_i\}_{i=1}^n$ that are explored by the data. In this setting, the integral kernel G in* (3.4) *becomes a positive semi-definite matrix in $\mathbb{R}^n$, so does $\overline{G}$ in* (3.5)*. Now the integral operator $\mathcal{L}_{\overline{G}}$ is defined by the matrix $\overline{G}$ on the weighted vector space $\mathbb{R}^n$ and its eigenvalues are the generalized eigenvalues of (G, B) with $B = \mathrm{Diag}(\rho(r_1), \cdots, \rho(r_n))$. As a result, the SIDA-RKHS H_G is the vector space spanned by the eigenvectors with nonzero eigenvalues. Furthermore, the norms in* (3.8) *can be computed directly from the eigen-decomposition. Viewing them as piecewise constant approximations of functions, these discrete approximation can be viewed as a special implementation of the numerical algorithm in Chapter* 3.2*.*

The next theorem characterizes the function space of identifiability. Furthermore, it shows that this inverse problem is ill-posed since the estimator requires the inverse of a compact operator.

Theorem 3.5 (Function space of identifiabilty)**.** *Suppose that Assumption* 3.1 *holds. Let* $\phi_N^v \in L^2(\rho)$ *be the Riesz representation of the bounded linear functional:*

$$\langle \phi_N^v, \psi \rangle_{L^2(\rho)} = \frac{1}{N} \sum_{k=1}^{N} \int R_\psi[u_k](x) v_k(x) dx, \, \forall \psi \in L^2(\rho). \tag{3.10}$$

Then the following statements hold.

(a) The Fréchet derivative of $\mathcal{E}(\phi)$ *in* $L^2(\rho)$ *is* $\nabla \mathcal{E}(\phi) = 2(\mathcal{L}_{\overline{G}}\phi - \phi_N^v)$.

(b) The function space of identifiability is $H = \overline{\text{span}\{\psi_i\}}$ *with closure in* $L^2(\rho)$*, where* $\{\psi_i\}$ *are eigenfunctions of* $\mathcal{L}_{\overline{G}}$ *with positive eigenvalues. Furthermore, the minimizer of* $\mathcal{E}(\phi)$ *in* H *is* $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^v$ *if* $\phi_N^v \in \mathcal{L}_{\overline{G}}(L^2(\rho))$*. In particular, if the data is perfect and generated from a true function* ϕ_{true}*, we have* $\phi_N^v = \mathcal{L}_{\overline{G}}\phi_{true}$ *and* $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^v = \phi_{true}$.

(c) The Fréchet derivative of $\mathcal{E}$ *in* H_G *is* $\nabla^{H_G}\mathcal{E}(\phi) = 2({\mathcal{L}_{\overline{G}}}^2\phi - \mathcal{L}_{\overline{G}}\phi_N^v)$*. Its zero leads to another estimator* $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-2}\mathcal{L}_{\overline{G}}\phi_N^v$ *if* $\phi_N^v \in \mathcal{L}_{\overline{G}}(L^2(\rho))$.

Proof of Theorem 3.5. From (3.7), we can write the loss functional in (3.1) as

$$\mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}}\phi, \phi \rangle_{L^2(\rho)} - 2\langle \phi_N^v, \phi \rangle_{L^2(\rho)} + C_N^v.$$

Then we can compute the Fréchet derivative directly from definition and Part (a) follows.

For Part (b), first note that for any $\phi_N^v \in \mathcal{L}_{\overline{G}}(L^2(\rho))$, the estimator $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^v$ is the unique zero of the loss functional's Fréchet derivative in H, hence it is the unique minimizer of $\mathcal{E}(\phi)$ in H. In particular, when the perfect data is generated from ϕ_{true}, i.e. $R_{\phi_{true}}[u_k] = v_k$, by (3.7) and

the definition of the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in (2.6), we have

$$\langle \phi_N^v, \psi \rangle_{L^2(\rho)} = \langle \mathcal{L}_{\overline{G}} \phi_{true}, \psi \rangle_{L^2(\rho)}$$

for any $\psi \in L^2(\rho)$. Thus, $\phi_N^v = \mathcal{L}_{\overline{G}} \phi_{true}$ and $\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1} \phi_N^v = \phi_{true}$. That is, $\phi_{true} \in H$ is the unique minimizer of the loss functional $\mathcal{E}$ for perfect data. Meanwhile, note that H is the orthogonal complement of the null space of $\mathcal{L}_{\overline{G}}$, and $\mathcal{E}(\phi_{true} + \phi^0) = \mathcal{E}(\phi_{true})$ for any ϕ^0 such that $\mathcal{L}_{\overline{G}} \phi^0 = 0$. Thus, H is the largest such function space, and we conclude that H is the function space of identifiability.

To prove Part (c), we further re-write the loss functional as

$$\mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}} \phi, \mathcal{L}_{\overline{G}} \phi \rangle_{H_G} - 2 \langle \mathcal{L}_{\overline{G}}{}^{1/2} \phi_N^v, \mathcal{L}_{\overline{G}}{}^{1/2} \phi \rangle_{H_G} + C_N^v,$$

which follows from (3.9) and the definition of $\langle \cdot, \cdot \rangle_{H_G}$. Thus, by definition, the Fréchet derivative of $\mathcal{E}(\phi)$ in the direction of $\psi \in H_G$ is

$$\begin{aligned} \langle \nabla^{H_G} \mathcal{E}(\phi), \psi \rangle_{H_G} &= \lim_{\epsilon \to 0} \frac{1}{\epsilon} [\mathcal{E}(\phi + \epsilon \psi) - \mathcal{E}(\phi)] \\ &= 2 \langle \mathcal{L}_{\overline{G}} \phi, \mathcal{L}_{\overline{G}} \psi \rangle_{H_G} - 2 \langle \mathcal{L}_{\overline{G}}{}^{1/2} \phi_N^v, \mathcal{L}_{\overline{G}}{}^{1/2} \psi \rangle_{H_G} \\ &= 2 \langle \mathcal{L}_{\overline{G}}{}^2 \phi - \mathcal{L}_{\overline{G}} \phi_N^v, \psi \rangle_{H_G}, \end{aligned}$$

which gives the Fréchet derivative $\nabla^{H_G} \mathcal{E}(\phi)$. □

Remark 3.6 (Regularization with the L^2 and the SIDA-RKHS norms)**.** *In practice, due to the discrete and/or noisy data, we often have $\phi_N^v = \mathcal{L}_{\overline{G}} \phi_{true} + \phi_1^\delta + \phi_2^\delta$, where the perturbation from the true function is decomposed to $\phi_1^\delta \in \mathcal{L}_{\overline{G}}(L^2(\rho))$ and $\phi_2^\delta \in \mathcal{L}_{\overline{G}}(L^2(\rho))^\perp$. Clearly, when $\phi_2^\delta \neq 0$,*

the estimator $\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1}\phi_N^v$ *does not exist and regularization is necessary. Next, we compare the* L^2 *and the SIDA-RKHS regularizers, i.e., consider the regularized loss functional with* $\mathcal{R}(\phi)$ *being* $\lambda\|\phi\|_{L^2}^2$ *and* $\lambda\|\phi\|_{H_G}^2$*. Then, their minimizers are*

$$\widehat{\phi}_\lambda^{L^2} = (\mathcal{L}_{\overline{G}} + \lambda I)^{-1}\phi_N^v, \quad \widehat{\phi}_\lambda^{H_G} = (\mathcal{L}_{\overline{G}}{}^2 + \lambda I)^{-1}\mathcal{L}_{\overline{G}}\phi_N^v.$$

Plugging in $\phi_N^v = \mathcal{L}_{\overline{G}}\phi_{true} + \phi_1^\delta + \phi_2^\delta$*, we have*

$$\begin{aligned}\widehat{\phi}_\lambda^{L^2} &= \phi_{true} + (\mathcal{L}_{\overline{G}} + \lambda I)^{-1}(\phi_1^\delta - \lambda\phi_{true} + \phi_2^\delta),\\ \widehat{\phi}_\lambda^{H_G} &= \phi_{true} + (\mathcal{L}_{\overline{G}}{}^2 + \lambda I)^{-1}(\mathcal{L}_{\overline{G}}\phi_1^\delta - \lambda\phi_{true}).\end{aligned}$$

A regularizer then selects the optimal λ *to balance the errors,*

$$\begin{aligned}\|\widehat{\phi}_\lambda^{L^2} - \phi_{true}\|_{L^2(\rho)}^2 &= \|(\mathcal{L}_{\overline{G}} + \lambda I)^{-1}(\phi_1^\delta + \phi_2^\delta)\|^2 + \|(\mathcal{L}_{\overline{G}} + \lambda I)^{-1}\lambda\phi_{true}\|^2,\\ \|\widehat{\phi}_\lambda^{H_G} - \phi_{true}\|_{L^2(\rho)}^2 &= \|(\mathcal{L}_{\overline{G}}{}^2 + \lambda I)^{-1}\mathcal{L}_{\overline{G}}\phi_1^\delta\|^2 + \|(\mathcal{L}_{\overline{G}}{}^2 + \lambda I)^{-1}\lambda\phi_{true}\|^2,\end{aligned}$$

where the first term on the right hand side requires a large λ*, whereas the second term requires a small* λ*. Assuming that the errors are much smaller than the true signal* ϕ_{true}*, the optimal* λ *would be small so that the second error is negligible. In this case, the bias in* $\widehat{\phi}_\lambda^{L^2}$ *is about* $\mathcal{L}_{\overline{G}}{}^{-1}(\phi_1^\delta) + \lambda^{-1}\phi_2^\delta$*, whereas the bias in* $\widehat{\phi}_\lambda^{H_G}$ *is about* $\mathcal{L}_{\overline{G}}{}^{-1}(\phi_1^\delta)$*. Thus, when* λ *is small, the SIDA-RKHS regularized estimator* $\widehat{\phi}_\lambda^{H_G}$ *is more accurate than the* L^2 *regularized estimator. To avoid amplifying the error* ϕ_2^δ*, a projection is necessary for the* L^2 *regularizer, and we will compare the projected* L^2 *regularizer with the SIDA-RKHS regularizer at the end of Chapter* 3.2*.*

2. Algorithm: nonparametric regression with DARTR

Algorithm: nonparametric regression with DARTR. Based on the identifiability theory in Section 3.1, we introduce next a nonparametric learning algorithm with Data Adaptive RKHS Tikhonov Regularization (DARTR). We briefly sketch the algorithm in the following four steps, whose details are presented in Appendix B.1.

1. Estimate the exploration measure ρ defined (3.2). We utilize the data to estimate the support of the true kernel and the exploration measure ρ. The support of the true kernel lies in $[0, d(\Omega)]$ with $d(\Omega)$ being the diameter of the domain Ω, and it is further confined from a combination between the support of v_k and the support of $g[u_k](x, y)$. Then, we constrain the discrete approximation of ρ on the support of ϕ.

2. Assemble the regression matrices and vectors, which will be repeatedly used. We select a class of hypothesis spaces $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ with basis functions $\{\phi_i\}$ and with dimension n in a proper range. Then, we compute the regression normal matrices and vectors, as well as the basis matrix,

$$\overline{A}_n(i,j) = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle, \quad \overline{b}_n(i) = \langle \phi_N^v, \phi_i \rangle_{L^2(\rho)}, \quad B_n(i,j) = \langle \phi_i, \phi_j \rangle_{L^2(\rho)}. \tag{3.11}$$

   from data for each of these hypothesis spaces.

3. Find the best regularized estimator $\widehat{c}_{\lambda_n}$ and the corresponding loss value $\mathcal{E}(\widehat{c}_{\lambda_n})$ by DARTR in Algorithm 1 for each triplet $(\overline{A}_n, \overline{b}_n, B_n)$.

4. Select the one with the smallest loss value $\mathcal{E}(\widehat{c}_{\lambda_n})$ from the estimators $\{\widehat{c}_{\lambda_n}\}_n$,.

Input: The regression triplet $(\overline{A}, \overline{b}, B)$ consisting of normal matrix $\overline{A}$, vector $\overline{b}$ and basis matrix B as in (3.11).

Output: SIDA-RKHS regularized estimator $\widehat{c}_{\lambda_0}$ and loss value $\mathcal{E}(\widehat{c}_{\lambda_0})$.

1: Solve the generalized eigenvalue problem $\overline{A}Q = BQ\Lambda$, where Λ is the diagonal matrix of eigenvalues and the matrix Q has columns being eigenvectors orthonormal in the sense that $Q^\top BQ = I$.

2: Compute the RKHS-norm matrix $B_{rkhs} = (Q\Lambda Q^\top)^{-1}$, using pseudo inverse when Λ is singular. We refer to Appendix B.1 Remark B.1 on a computational technique to avoid the inverse matrix.

3: Use the L-curve method to find an optimal estimator $\widehat{c}_{\lambda_0}$: select λ_0 maximizing the curvature of the λ-curve $(\log \mathcal{E}(\widehat{c}_\lambda), \log(\widehat{c}_\lambda^\top B_{rkhs}\widehat{c}_\lambda))$, where the least squares estimator $\widehat{c}_\lambda = (\overline{A}+\lambda B_{rkhs})^{-1}\overline{b}$ minimizes the regularized loss function

$$\mathcal{E}_\lambda(c) = \mathcal{E}(c) + \lambda c^\top B_{rkhs} c \;\text{ with }\; \mathcal{E}(c) = c^\top \overline{A} c - 2c^\top \overline{b} + \overline{b}^\top \overline{A}^{-1}\overline{b},$$

where the matrix inversion is a pseudo-inverse when it is singular.

Algorithm 1: Data Adaptive RKHS Regularization (DARTR).

In comparison to the classical nonparametric regression using only $(\overline{A}_n, \overline{b}_n)$, the novelty of our algorithm is the data adaptive components, such as the exploration measure ρ, the basis matrix B_n in $L^2(\rho)$ and the norm of the SIDA-RKHS for regularization. The computation of the SIDA-RKHS norm is based on the generalized eigenvalues problem with the pair $(\overline{A}_n, B_n)$, whose eigenvalues approximate the eigenvalues of $\mathcal{L}_{\overline{G}}$ in (3.6) and $\widehat{\psi}_k = Q_{jk}\phi_j$ approximates the eigenfunctions of $\mathcal{L}_{\overline{G}}$ (see Theorem 3.7). The additional computational cost is only the generalized eigenvalue problem, which can be solved efficiently.

Theorem 3.7. *Let $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n \subset L^2(\rho)$ and let $(\overline{A}_n, B_n)$ be the normal and basis matrix in (3.11). Assume that $\mathcal{H}_n$ is large enough so that $\mathcal{L}_{\overline{G}}(L^2(\rho)) \subset \mathcal{H}_n$ (which is true, for example*

when ρ is a discrete-measure on a discrete set $\mathcal{S}$ and $\{\phi_n\}$ are piecewise constant functions with $n = |\mathcal{S}|$). Then, the operator $\mathcal{L}_{\overline{G}}$ in (3.6) *has eigenvalues $(\lambda_1, \ldots, \lambda_n)$ solved by the generalize eigenvalue problem*

$$\overline{A}_n Q = B_n \Lambda Q, \quad s.t., Q^\top B_n Q = I_n, \quad \Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n), \tag{3.12}$$

and the corresponding eigenfunctions of $\mathcal{L}_{\overline{G}}$ are $\{\psi_k = \sum_{j=1}^n Q_{jk}\phi_j\}$.

Proof. Let $\psi_k = \sum_{j=1}^n Q_{jk}\phi_j$ with $Q^\top B_n Q = I_n$. Then, ψ_k is an eigenfunction of $\mathcal{L}_{\overline{G}}$ with eigenvalue λ_k if and only if for each i,

$$\langle \phi_i, \lambda_k \psi_k \rangle_{L^2(\rho)} = \langle \phi_i, \mathcal{L}_{\overline{G}} \psi_k \rangle_{L^2(\rho)} = \sum_{j=1}^n \langle \phi_i, \mathcal{L}_{\overline{G}} \phi_j \rangle_{L^2(\rho)} Q_{jk} = \sum_{j=1}^n \overline{A}_n(i,j) Q_{jk},$$

where the last equality follows from the definition of $\overline{A}_n$ in (3.11). Meanwhile, by the definition of B_n we have $\langle \phi_i, \lambda_k \psi_k \rangle_{L^2(\rho)} = \sum_{j=1}^n B_n(i,j)\lambda_k Q_{jk}$ for each i. Then, Equation (3.12) follows. □

Comparison with projected l^2 and L^2 regularizers. Our DARTR method differs from other regularizers in its use of the SIDA-RKHS norm, which restricts the function to be in the function space of identifiability. In the following, we compare it with the l^2 and L^2 regularizers that apply regularization terms $\mathcal{R}(\phi) = \|\phi\|_{l^2}^2 = \sum_i c_i^2$ or $\mathcal{R}(\phi) = \|\phi\|_{L^2}^2 = c^\top B_n c$. In fact, a direct application of these two regularization terms would lead to problematic regularizers with largely biased estimators when $\overline{A}_n$ is singular, i.e., when the function space of identifiability is a proper subspace of $L^2(\rho)$, because the inverse problem is ill-defined on $L^2(\rho)$. Thus, in practice, one makes a projection to the function space of identifiability (i.e., the image of $\overline{A}_n$ in computation) before adding these regularization terms, and we call them projected l^2 and L^2 regularizers.

Table 3.1: The SIDA-RKHS regularizer v.s. the projected l^2, L^2 regularizers.

	l^2	L^2	SIDA-RKHS
$\mathcal{R}(\phi)$	$\|c\|^2 = c^\top c$	$\|c\|^2_{B_n} = c^\top B_n c$	$\|c\|^2_{H_G} = c^\top B_{rkhs} c$
c_λ	$c_\lambda = \sum_{i=1}^{l} \frac{1}{\sigma_i+\lambda} p_i^\top \overline{b}_n$	$c_\lambda = \sum_{i=1}^{l} \frac{1}{\lambda_i+\lambda} q_i^\top \overline{b}_n$	$c_\lambda = \sum_{i=1}^{l} \frac{1}{\lambda_i+\lambda\lambda_i^{-1}} q_i^\top \overline{b}_n$
SVD	$\overline{A}_n = \sum_{i=1}^{n} \sigma_i p_i p_i^\top, p_i^\top p_j = \delta_{ij}$ $P^\top \overline{A}_n P = \Sigma$, $P^\top P = I$	$\overline{A}_n = \sum_{i=1}^{n} \lambda_i q_i q_i^\top, \;\; q_i^\top B_n q_j = \delta_{ij}$ $Q^\top \overline{A}_n Q = \Lambda$, $Q^\top B_n Q = I$	

All regularizers estimate $\phi = \sum_{i=1}^{n} c_i \phi_i$ from $\overline{A}_n c = \overline{b}_n$ with basis matrix B_n (see (3.11)). The projected l^2 and L^2 regularizers use only the non-zero eigenvalues $\{\sigma_i\}_{i=1}^{l}$ and $\{\lambda_i\}_{i=1}^{l}$ and their eigenvectors. Thus, the projected l^2 and L^2 regularizers involve an additional step of projection to the function space of identifiability, before adding the regularization term $\mathcal{R}(\phi)$ to the loss function.

Table 3.1 compares our SIDA-RKHS regularizer with the projected l^2 and L^2 regularizers. We note that there are the following connections:

- The L^2 regularizer is a basis-adaptive generalization of the l^2 regularizer. When $B_n = I$ (i.e., the basis $\{\phi_i\}$ are orthonormal in $L^2(\rho)$), the two are the same. When B_n is not the identity matrix (i.e., the basis $\{\phi_i\}$ are not orthonormal in $L^2(\rho)$), which happens often, the L^2 regularizer takes it into account through the generalized eigenvalue problem.
- The SIDA-RKHS regularizer is an improvement over the L^2 regularizer. When all the generalized eigenvalues are $\lambda_i \equiv 1$ (e.g., $\mathcal{L}_{\overline{G}}$ is an identity operator), the two are the same.
- The SIDA-RKHS regularizer restricts the learning to be in the function space of identifiability by definition, while the other two regularizers may miss this important restriction.

3. Numerical tests on synthetic data

We test our learning method on three types of operators: linear integral operators, nonlocal operators and nonlinear operators. For each type of operators, we systematically examine the method in the regimes of noiseless and noisy data, with kernels in and out of the SIDA-RKHSs. Since the ground-truth kernel is known, we study the convergence of estimators to the true kernel as the data mesh refines. Thus, the regularization has to overcome both numerical error and noise in the imperfect data. All codes used will be publicly released on GitHub.

Comparison with baseline methods. On each dataset, we compare our SIDA-RKHS regularizer with two baseline regularizers using the projected l^2 and L^2 regularizers (denoted by l2 and L2 in the figures below, respectively) defined in Table 3.1. All three regularizers use the same L-curve method to select the hyper-parameter λ as described in Appendix B.2. They differ only at the regularization norm. In numerical implementation with $\phi = \sum_{i=1}^n c_i \phi_i$, in a hypothesis space $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$, the $\mathcal{R}(\phi)$ terms for l^2, L^2 and SIDA-RKHS are computed as $\|\phi\|_{l^2}^2 = \sum_i c_i^2$, $\|\phi\|_{L^2}^2 = c^\top B_n c$ and $\|\phi\|_{H_G}^2 = c^\top B_{rkhs} c$ with $B_n = \text{diag}(\rho)$ and B_{rkhs} specified in the DARTR Algorithm 1.

Settings of synthetic data. We test two kernels for each type of operators:

• *Truncated sine kernel.* The truncated sine kernel is $\phi_{true}(x) = \sin(2x)\mathbf{1}_{[0,3]}(x)$. It represents a kernel with discontinuity. Due to the nonlocal dependency of the operator on the kernel, this discontinuity can cause global bias of the estimator in the inverse problem.

• *Gaussian kernel.* The kernel ϕ_{true} is the Gaussian density centered at 3 with standard deviation 0.75. It represents a smooth kernel whose interaction concentrated in the middle of its support.

The kernels act on the same set of function $\{u_k\}_{k=1,2}$ with $u_1 = \sin(x)\mathbf{1}_{[-\pi,\pi]}(x)$ and $u_2(x) =$

$\sin(2x)\mathbf{1}_{[-\pi,\pi]}(x)$. When generating the data for learning, the integral $R_\phi[u_k] = v_k$ is computed by the adaptive Gauss-Kronrod quadrature method. This integrator is much more accurate than the Riemann sum integrator that we will use in the learning stage. To create discrete datasets with different resolutions, for each $\Delta x \in 0.0125 \times \{1, 2, 4, 8, 16\}$, we take values of $\{u_k, v_k\}_{k=1}^N = \{u_k(x_j), v_k(x_j) : x_j \in [-40, 40], j = 1, \ldots, J\}_{k=1}^N$, where x_j is a point on the uniform grid with mesh size Δx. For the nonlinear operator, to avoid the inverse problem being ill-defined, we introduce add an additional pair of data (u_3, v_3) with $u_3(x) = x\mathbf{1}_{[-\pi,\pi]}(x)$ (see Section 3.2, Chapter 2 for more details). In short, the discrete data $\{u_k\}_{k=1,2}$ are continuous functions and the discrete data u_3 is a piece-wise continuous function.

For each kernel, we consider both noiseless and noisy data with different noise levels by taking values of noise-to-signal-ratio (nsr) in $\{0, 0.5, 1, 1.5, 2\}$. Here the noise is added to each spatial mesh point, independent and identically distributed centered Gaussian with standard deviation σ, and the noise-to-signal-ratio is the ratio between σ and the average L^2 norm of v_k.

Settings for the learning algorithm. When estimating the kernels from the discrete data, we estimate the values of the kernel on the points $\mathcal{S} = \{r_j\}_{j=1}^J$ with $r_j = j\Delta x$, the support of the empirical exploration measure ρ. When the data mesh refines, the size of this set increases. In view of the algorithm in Chapter 3.2, such a discrete estimation uses a hypothesis space with B-spline basis functions consisting of piece-wise constants with knots being the points in $\mathcal{S}$. Thus, this hypothesis space has the largest dimension with the basis matrix B_n in (3.11) being non-singular, and there is no need to select the optimal dimension. In this setting, the regularizer is the only source of regularization and there is no regularization from basis functions. Hence, this setting highlights the role of the Tikhonov-type regularizers.

Performance assessment. We assess the performance of the regularizers by their ability to

consistently identify the true kernels in the presence of numerical error (in the Riemann sum approximation of the integrals due to discrete data) and noise (due to noisy data). We present typical estimators, the $L^2(\rho)$ errors of the estimators as data mesh refines, as well as the statistics (mean and standard deviation) of the rates of convergence that are computed from 20 independent simulations.

Summary of main results. Our main finding are as follows.

- The SIDA-RKHS regularizer's estimators are more accurate than those of the l^2 and L^2 regularizers when the regularization parameter is properly selected. However, multiple factors are involved in the selection of the parameter, ranging from the form of the operator, the numerical approximation, the noise and the treatment of the singular or ill-conditioned normal matrix. Thus, in addition to accuracy of the estimator, it is also important to also compare the consistency of convergence rates for different levels of noise.
- The SIDA-RKHS regularizer robustly leads to estimators converging at a consistent rate for all levels of noises for each operator, while the other two regularizers cannot.
- The rate of convergence of the SIDA-RKHS regularizer's estimator from noisy data depends on both the continuity of the kernel and the continuity of the discrete data: when the two matches, the rate is close to 1; otherwise, the rate can are lower than 1, as shown in Table 3.2.

3.1. Linear integral operators We consider first the integral operator with kernel ϕ:

$$R_\phi[u](x) = \int_\Omega \phi(|y-x|)u(y)dy. \tag{3.13}$$

Table 3.2: Rate of convergence of the SIDA-RKHS regularizer's estimators from noisy data.

Kernel	Linear Integral Operator Data continuity(C)	Nonlinear Operator Data continuity (D)	Nonlocal Operator Data continuity (C)
Truncated Sine (D)	0.29	**0.94**	0.29
Gaussian (C)	**0.62**	0.66	**1.01**

* Here "C" stands for continuous, and "D" stands for discontinuous. When the continuity of the kernel and data matches, the rates are close to 1. The rates are the average of the mean rates for nsr $\in \{0.1, 0.5, 1, 2\}$ in the right columns of Figure 3.1-3.3. We do not report the rate for the l^2 and L^2 regularizers because they do not have a consistent rate.

After a change of variables in the integral, it is the operator R_ϕ in (2.1) with $g[u](x, y) = u(x + y)$. Such kernels in operators arise in a wide range of applications, such as the Green's function in PDEs [24, 31] and convolution kernels in image processing [35], to name just a few.

For this operator, the exploration measure ρ (defined in (3.2)) is a uniform measure, since each data $g[u_k]$ interacts with the kernel uniformly. Furthermore, since each $g[u_k]$ is continuous, the reproducing kernel $\overline{G}$ in (3.4) is continuous on the support of ρ, thus the SIDA-RKHS consists of continuous functions. As a result, we expect the algorithm to learn the smooth Gaussian kernel better than the discontinuous truncated sine kernel.

The left column of Figure 3.1 shows the typical estimators by the three regularizers, in comparison of the true kernel, when $\Delta x = 0.05$ and noise-to-signal-ratio nsr = 1. The exploration measure ρ (in cyan color) is uniform for each kernel, and its support, estimated from the difference between the supports of $g[u_k]$ and v_k, is slightly larger than the support of the true kernel. All three regularizers lead to accurate estimators. The RKHS regularizer's estimators are closest to the true kernel and this is further verified in the middle 3-column panel with $\Delta x = 0.05$ add nsr = 1: for the truncated sine kernel, all three estimators' $L^2(\rho)$ errors are about 10^{-1}; but for the Gaussian kernel, the RKHS's estimator has an error close to 10^{-3} while the other two regularizers' error are about 10^{-2}.

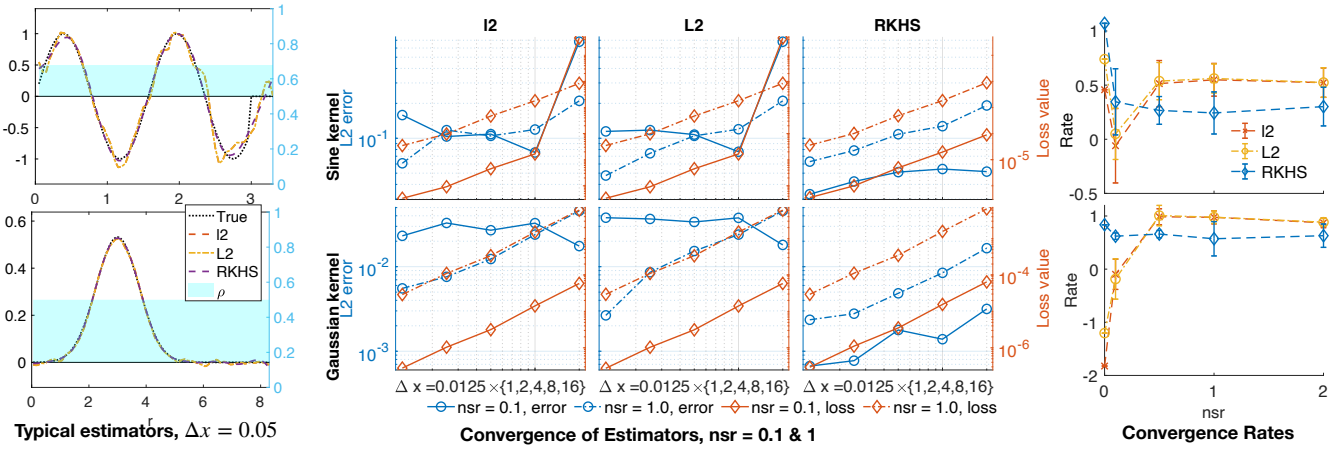


Figure 3.1: Linear integral operators with the truncated sine kernel and Gaussian kernel. Left Column: typical estimators by the three regularizers, in comparison of the true kernel, superimposed with the exploration measure ρ (in cyan color), when $\Delta x = 0.05$ and noise-to-signal-ratio nsr $= 1$. Middle 3-Columns: convergence of estimators as the data mesh-size Δx refines, along with values of the loss function. Right Column: the mean and standard deviation of the convergence rates in 20 independent simulations, with five levels of noise (with nsr $\in \{0, 0.1, 0.5, 1, 2\}$). Only the SIDA-RKHS regularizer's estimator consistently converges for all levels of noise, and its estimators are mostly more accurate than those of the other two regularizers.

The middle 3-column panel of Figure 3.1 shows the convergence of the estimator's $L^2(\rho)$ error as the data mesh refines when nsr $= 0.1$ and nsr $= 1$, superimposed with the corresponding values of the loss function. When nsr $= 1$, all three regularizers' estimators converge for both kernels, at rates that are close to the rates of the loss function, and their errors are comparable. However, when nsr $= 0.1$, the RKHS regularizer continues to yield converging estimators, whereas the other two regularizers have flat error lines even though the corresponding loss values keep decaying. In particular, those flat error lines are above those errors for nsr $= 1$ with $\Delta x \leq 0.025$, i.e., when the numerical error is small. Thus, these results demonstrate the importance to take into account the function space of learning via SIDA-RKHS, particularly when the noise level is relatively low.

The right column of Figure 3.1 shows the mean and standard deviations of the rates of convergence in 20 independent simulations. The RKHS regularizer has consistent rates of convergence

for all levels of noises. The rates are close to 1 for the smooth Gaussian kernel. The rates are below 0.5 for the discontinuous truncated sine kernel when the data is noisy and the rate is 1 when the data is noiseless. On the other hand, the l^2 and L^2 regularizers fails to have consistent rates when the noise level reduces. They present deceivingly higher rates than the RKHS regularizer when nsr $\in \{0.5, 1, 2\}$, and the middle 3-column panel reveals the facts: they often have much larger errors than the RKHS when $\Delta x = 0.2$, thus leading to deceiving better rates even when their errors remains large as Δx decreases.

In short, the RKHS regularizer leads to estimators that converge consistently, at lower rates for the discontinuous truncated sine kernel (which does not match the continuity of the data) and at higher rates for the smooth Gaussian kernel (which match the continuity of the data), while the l^2 and L^2 regularizers cannot. Furthermore, RKHS regularizer's estimators are often more accurate than those of the other two regularizers.

3.2. Nonlinear operators Next we consider the nonlinear operator R_ϕ with $g[u](x, y) = \partial_x[u(x+y)u(x)]$:

$$R_\phi[u](x) = \int_\Omega \phi(|y|)\partial_x[u(x+y)u(x)]dy = [u * \phi(|\cdot|)u]'(x). \tag{3.14}$$

Such nonlinear operators arise in the mean-field equations of interaction particles [41, 53, 66, 72], and the function ϕ is called an interaction kernel. More precisely, the mean-field equations are of the form $\partial_t u = \nu\Delta u + \mathrm{div}(u * K_\phi u)$ on $\mathbb{R}^d$, where $K_\phi(y) = \phi(|y|)\frac{y}{|y|}$. Here we consider only $d = 1$ and neglect the ratio $\frac{y}{|y|}$ to obtain the above operator.

We add an additional pair of data (u_3, v_3) with $u_3(x) = x\mathbf{1}_{[-\pi,\pi]}(x)$, so as to avoid the issue that the value of $[u * \phi(|\cdot|)u](x)$ is under-determined from the data $v(x) = [u * \phi(|\cdot|)u]'(x)$ due to the differential. Here we set the derivative of u_3 to be $u_3'(x) = \mathbf{1}_{[-\pi,\pi]}(x)$. These derivatives are

approximated by finite difference when learning the kernel from discrete data. Note that the u_3 and its derivative have jump discontinuities. As a result, the reproducing kernel $\overline{G}$ in (3.4) also has discontinuity, and the SIDA-RKHS contains discontinuous functions.

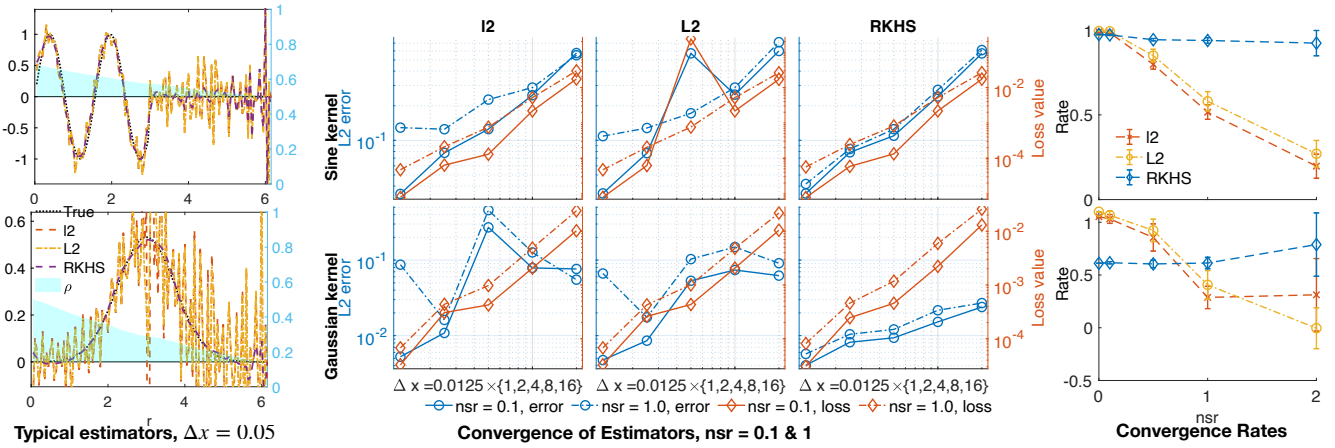


Figure 3.2: Nonlinear operators with the truncated sine kernel and Gaussian kernel. The SIDA-RKHS regularizer's estimators are significantly more accurate than those of the l^2 and L^2 regularizers in the left column. The middle 3-column panel shows that the SIDA-RKHS regularizer leads to consistently converging estimators as the data mesh refines, for both levels of noise, while the other two regularizers have slower and less consistent error decay and their error lines flatten when the noise level is nsr = 1. The right column shows that only the SIDA-RKHS regularizer has consistent rates for all levels of noise, and the other two regularizers' rates drops significantly when the noise level increases.

The left column of Figure 3.2 shows that the exploration measure ρ is non-uniform due to the nonlinear function $g[u_k]$. Its density is a decreasing function, suggesting that the data explores the short range interactions more than the long range interaction. The RKHS regularizer's estimators significantly outperform theose of the other two regularizers, and they are near smooth and are close to the true kernels. The l^2 and L^2 regularizers lead to largely oscillating estimators, suggesting an overfitting. Note that the RKHS estimators also have oscillating parts, but they are only in the region where the exploration measure has little weight, due to the limited data exploration. The superior performance of RKHS regularizer is further verified in the middle 3-column panel with

$\Delta x = 0.05$ add nsr $= 1$: its errors are much smaller than those of the other two regularizers.

The middle 3-column panel of Figure 3.2 shows that the RKHS regularizer's error consistently decreases as the data mesh refines. In contrast, the other two regularizers have slower and less consistent error decay, in particular, their error lines flatten as the noise level increases.

The right column of Figure 3.2 shows that the RKHS regularizer has consistent rates of convergence for all levels of noises, with all rates close to 1 for the truncated sine kernel, and slightly above 0.5 for the Gaussian kernel. In comparison, the other two regularizers' rates decrease as the noise level increases, dropping to close zero when the noise level is nsr $= 2$.

In short, the RKHS regularizer's estimators are more accurate than those of the l^2 and L^2 regularizers. More importantly, the RKHS regularizer consistently leads to convergent estimators, maintaining similar rates for all levels of noises, at rates close to 1 for the truncated sine kernel (which is discontinuous, matching the discontinuity of data) and at rates slightly above 0.5 for the Gaussian kernel (which is smooth, not matching the discontinuity of data). The l^2 and L^2 regularizers have convergent estimators, but the rates of convergence drop when the noise level increases.

3.3. Nonlocal operators At last, we consider nonlocal operators R_ϕ with $g[u](x, y) = u(x+y) - u(x)$:

$$R_\phi[u](x) = \int_\Omega \phi(|y|)[u(x + y) - u(x)]dy. \tag{3.15}$$

Such nonlocal operators arise in various areas such as nonlocal and fractional diffusions [3, 12, 23]. It have been used to construct homogenized models for peridynamic [95, 96].

The left column of 3.3 shows typical estimators. The exploration measure ρ shrinks to zero near the origin due to the difference $g[u] = u(y) - u(x)$ and the continuity of u. All three regularizers

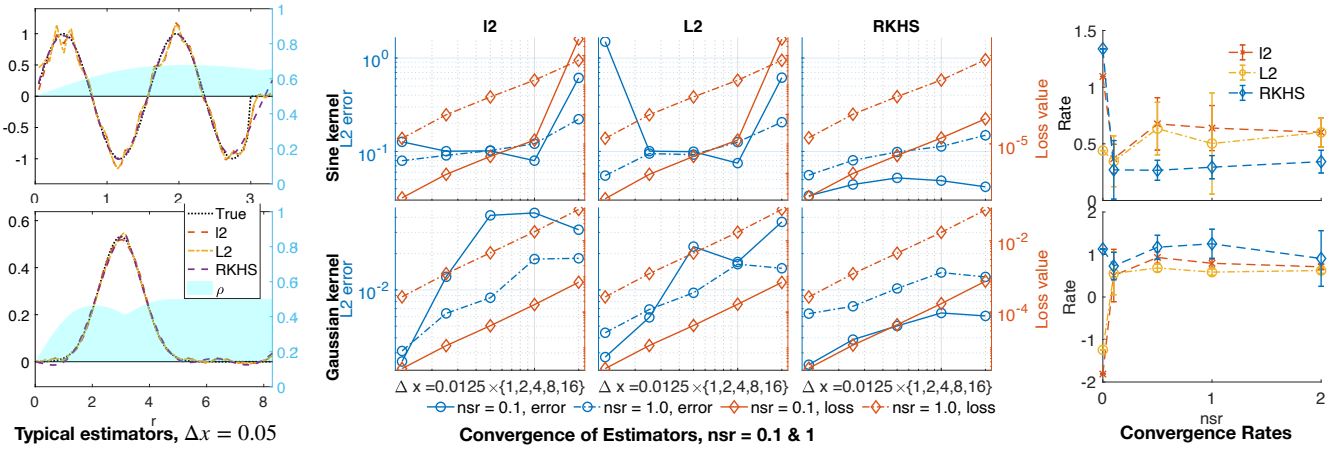


Figure 3.3: Nonlocal operators with the truncated sine kernel and Gaussian kernel. The left column shows that all regularizers lead to accurate estimators. The middle 3-column panel shows that the SIDA-RKHS regularizer leads to converging estimators as the data mesh refines for two levels of noise, though at a slow rate for the truncated sine kernel. The l^2 and L^2 regularizers have less consistent error decay for different noise levels and different kernels. Overall, the SIDA-RKHS estimators have smallest error mostly. The right column shows that only the SIDA-RKHS regularizer has consistent rates for all levels of noise, and the other two regularizers' rates drop significantly when the noise level increases.

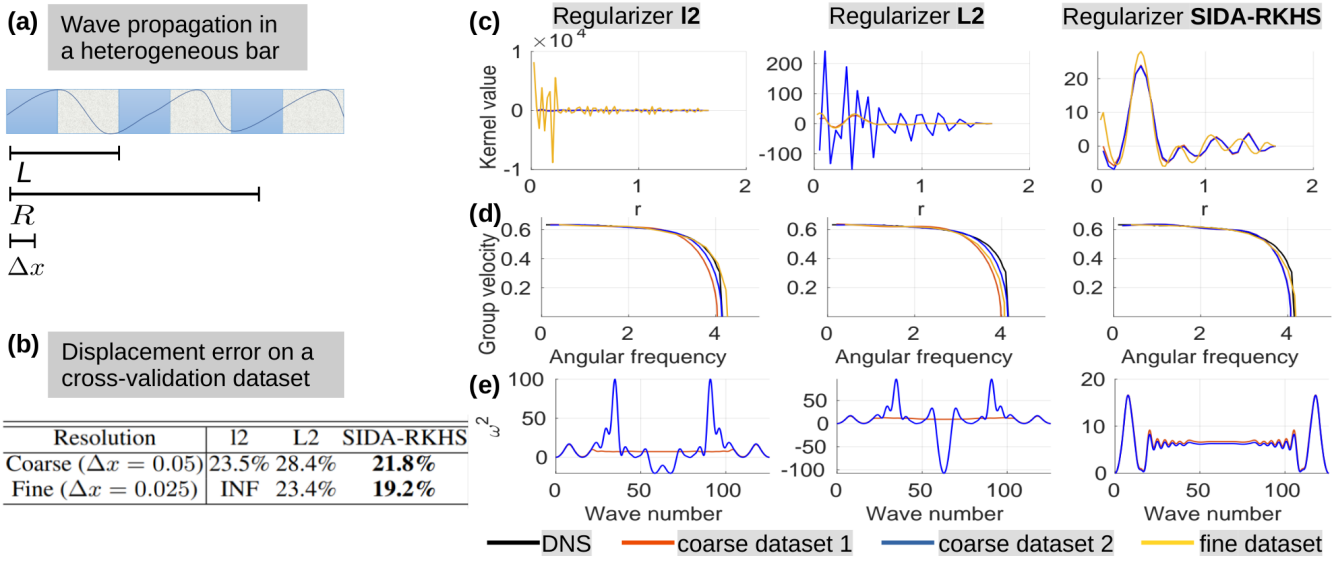
lead to accurate estimators, and the RKHS estimator is sightly closer to the true truncated sine kernel than the other two estimators.

In the middle 3-column panel of 3.3, we observe again that the RKHS regularizer leads to estimators converging as data mesh refines for both noise levels, even though the errors decay slower than the loss function values. On the other hand, the l^2 and L^2 regularizers have inconsistent error decay: the errors decreasing monotonically when nsr $= 1$, but the error lines oscillate when nsr $= 0.1$ for the truncated sine kernel.

The right column of 3.3 further confirms the consistency of the RKHS regularizer's rates and the inconsistency of the l^2 and L^2-regularizers' rates. When the data is noisy, the rates of the RKHS regularizer are about 0.29 for the truncated sine kernel (which has a jump discontinuity, not matching the continuity of the data) and about 1 for the Gaussian kernel (which is continuous,

matching the continuity of the data). Meanwhile, the rates for the l^2 and L^2-regularizers are about 0.65 for the truncated sine kernel, and about 0.8 for the Gaussian kernel. We note again that their better rates for the truncated sine kernel occur when their errors are larger than those of the RKHS estimators. Moreover, when the data is noiseless, RKHS regularizer has rates close to 1 for both kernels as desired, while the other two regularizers rates are not consistent.

4. Homogenization of wave propagation in meta-material



Resolution	l2	L2	SIDA-RKHS
Coarse ($\Delta x = 0.05$)	23.5%	28.4%	**21.8%**
Fine ($\Delta x = 0.025$)	INF	23.4%	**19.2%**

Figure 3.4: Wave propagation in a heterogeneous bar with ordered microstructure.

We seek a nonlocal homogenized model for the stress wave propagation in a one-dimensional heterogeneous bar with a periodic microstructure. For this problem, the goal is to obtain an effective surrogate model from high-fidelity (HF) datasets generated by solving classical wave equation, acting at a much larger scale than the size of the microstructure. Differing from previous examples, this problem has no ground-truth kernel. Therefore, we evaluate the estimator by

measuring its effectiveness of reproducing HF data in applications that are subject to different loading conditions with a much longer time from the problems used as training data.

For both training and validation purposes we use the HF dataset generated by the direct numerical solver (DNS) introduced in [83], which provides exact solutions of velocities including the appropriate jump conditions for the discontinuities in stress that occur at waves. Although the DNS has high accuracy on wave velocity, it is not suitable for long-term prediction because it requires the modeling of wave propagation through thousands of microstructural interfaces, which makes the computational cost prohibitive. To accelerate the computation, we approximate the HF model by a nonlocal model:

$$\partial_{tt}u(x,t) - R_\phi[u](x,t) = g(x,t), \text{ for } (x,t) \in \Omega \times [0,T], \tag{3.16}$$

where R_ϕ can be viewed as a nonlocal operator in the form of (3.15) with a kernel ϕ.

Experiment settings. We consider four types of data: three for training and one for validation of our algorithm. Three types of training datasets are employed: In Type 1 dataset, the bar is subject to an oscillating source $g(x,t)$; In Type 2 dataset, a boundary velocity loading $\partial_t u(-50,t) = \cos(jt)$ is applied; In Type 3 dataset, all settings are the same as in Type 2, except that the $\cos(jt)$ type loading is replaced by $\sin(jt)$. In all training datasets, we consider a relatively small domain $\Omega = [-50, 50]$ and short time $t \in [0, 2]$. Two spatial resolutions, $\Delta x = 0.05$ and $\Delta x = 0.025$ are considered, which we denote as the "coarse" and "fine" datasets, respectively.

With these three types of training datasets, we design three experiment settings to validate our method:

- *Coarse dataset* 1: we train the estimator using "coarse" dataset of Types 1 and 2.
- *Coarse dataset* 2: we train the estimator using "coarse" dataset of Types 1 and 3. By comparing

the learnt estimator from this setting with the result from the setting above, we mean to investigate the sensitivity of the inverse problem with respect to the choice of datasets.

• *Fine dataset:* we train the estimator using "fine" dataset of Types 1 and 2. By comparing the learnt estimator from this setting with the result from setting 1, we aim to check the convergence of the estimator with increasing data resolution. Note that the problem might becomes more ill-posed when decreasing Δx. Therefore, proper regularization is expected to become more important.

Additionally, we create a validation dataset, denoted as Type 4 dataset, very different from the training dataset. It considers a much longer bar ($\Omega = [-133.3, 133.3]$), under a different loading condition from the training dataset, and with a 50 times longer simulation time ($t \in [0, 100]$). Therefore, the cross-validation error checks the generalizability of the estimators. Further details of experiment settings can be found in Appendix B.3.

Results assessment. We present the learnt estimators in Figure 3.4. Since there is no ground-truth kernel, we assess the performance of each estimator based on three criteria. Firstly, we report in Figure 3.4(b) the prediction L^2 error of displacement on the cross validation dataset at $T = 100$. Secondly, we report in Figure 3.4(d) the resultant estimators the group velocity curves from our model and compare them with the curves computed with DNS. These curves directly depicts how much our surrogate model reproduces the dispersion properties in the heterogeneous material. At last, the learnt model should provide a physically stable material model. To check this, we also report the dispersion curve in 3.4(e). Its positivity indicates that the learnt nonlocal model is physically stable.

Performance of the estimators. Comparing the three estimators in Figure 3.4(c), one can see that only the SIDA-RKHS regularizer obtains consistent estimators in all three experiment settings. The oscillatory estimators of regularizers with l^2 or L^2-norm verify the ill-posedness, and

highlight the importance of using proper regularizers in nonlocal operator learning methods. The dispersion curves in Figure 3.4(e) stress the importance of regularizer from another aspect of view: our SIDA-RKHS regularizer provides physically stable material models in all settings, while the regularizers with l^2 or L^2-norm may result in highly oscillatory and non-physical models.

We further examine the regularized estimator in terms of its capability in reproducing DNS simulations through the prediction error of u on the cross validation dataset. When $\Delta x = 0.025$, it takes about 48 hours for the DNS simulation to generate one sample, while the homogenized nonlocal model only requires less than 20 minutes. From Figure 3.4(b), we can see that when $\Delta x = 0.05$, all three regularizers are robust and able to reproduce the DNS simulation with a reasonable accuracy ($\sim 20\%$). When we increase the data resolution to $\Delta x = 0.025$, the estimated nonlocal model from l^2 regularizer becomes unstable, which again verifies our analysis: when the data mesh refines, the kernel learning problem becomes more ill-posed and a good regularizer becomes a necessity. Meanwhile, both the L^2 and SIDA-RKHS regularizers lead to a more accurate estimator, indicating a trend of convergence. On both datasets, the SIDA-RKHS regularizer obtains the most accurate estimators.

5. Discussion and future directions

We have proposed the data-adaptive RKHS Tikhonov regularization (DARTR) method for the nonparametric learning of kernel functions in operators. The DARTR method regularizes the least squares regression by the norm of a SIDA-RKHS, which constraints the learning to the function space of identifiability.

Our numerical tests on synthetic data suggests that DARTR has the following advantages: (1) it is naturally adaptive to both data and the operator; (2) it is robust to numerical error due

to discrete data and white noise in data, leading to estimators converging at a consistent rate for different levels of noises; (3) it is computationally as efficient as classical nonparametric least squares regression methods, requiring in addition only an exploration measure and a basis matrix that come with negligible computational cost.

This study presents a preliminary introduction of the DARTR method. There are several directions for further development and analysis of DARTR in general settings and applications:

1. Convergence analysis. We have obtained convergent regularized estimators, but a convergence analysis is left as future work. The main difficulty to overcome is the complex combination of three factors: operator spectrum decay, the errors from numerical integration and noise, and regularization.
2. Multivariate kernel functions. When the kernel is a multivariate function, sparse-grid representation or sparse basis functions (sparse polynomials) become necessary. The regression will face the well-known curse-of-dimensionality. A related issue is to select the optimal dimension of the hypothesis space. Our identifiability theory remains valid. Thus, a future direction is to utilize methods such as kernel-regression or neural networks and further develop the SIDA-RKHS regularization.
3. Applications to Bayesian inverse problems. In a Bayesian perspective, the Tikhonov regularization can be interpreted as a Gaussian prior with a covariance matrix corresponding to the penalty term. In this perspective, our SIDA-RKHS norm coincides with the Zellner's g-prior ([6, 99]) that uses $\overline{A}^{-1}$ as prior covariance, because when the data has additive white noise and when the basis functions are orthonormal in $L^2(\rho)$, we have $B_{rkhs} = \overline{A}^{-1}$.
4. The DARTR method is applicable to general linear inverse problems that minimize a quadratic loss functional.

Chapter 4

Unsupervised learning of observation functions in state space models

1. Introduction and related work

We consider the following state space model for processes (X_t, Y_t) in $\mathbb{R} \times \mathbb{R}$:

$$\text{State equation:} \qquad dX_t = a(X_t)dt + b(X_t)dB_t, \qquad \text{with } a, b \text{ are known;} \tag{4.1}$$

$$\text{Observation equation:} \qquad Y_t = f_*(X_t), \qquad \text{with } f_* \text{ unknown.} \tag{4.2}$$

Here B_t is the standard Brownian motion, the drift function $a(x)$ and the diffusion coefficient $b(x)$ are given, satisfying the linear growth and global Lipschitz conditions. We assume that the initial distribution of X_{t_0} is given. Thus, the distribution of the state process (X_t) is known.

Our goal is to estimate the unknown observation function f_* from data consisting of a large ensemble of trajectories of the process Y_t, denoted by $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$, where m indexes trajectories,

and $t_0 < \cdots < t_L$ are the times at which the observations are made. In particular, there are no pairs (X_t, Y_t) being observed, so in the language of machine learning this may be considered an unsupervised learning problem. A case of particular interest in the present work is when the observation function f_* is nonlinear *and non-invertible*. We will also emphasize the usefulness of many short trajectories (vs. few long trajectories), albeit both the theory and algorithms that we consider are generally applicable in a wide range of regimes.

We estimate the observation function f_* by matching generalized moments, while constraining the estimator to a suitably chosen finite-dimensional hypothesis space, whose dimension depends on the number of observations, in the spirit of nonparametric statistics. We consider both first and second moments, as well as temporal correlations, of the observation process. The estimator minimizes the discrepancy between the moments over an hypothesis space, with upper and lower pointwise constraints estimated from data. The method we propose has several significant strengths:

- the generalized moments do not require the invertibility of the observation function f_*;
- low-order generalized moments tend to be robust to additive observation noise;
- generalize moments avoid the need of local constructions, since they depend on the entire distribution of the latent and observed processes;
- our nonparametric approach does not require *a priori* information about the observation function, and it can deal with both regular and piecewise regular functions;
- the method is computationally efficient because the moments need to be estimated only once, and the computation is can performed in parallel.

We note that the method we propose readily extends to multivariate state space models, with the main statistical and computational bottlenecks coming from the curse of dimensionality in the representation and estimation of a higher-dimensional f_* in terms of basis functions.

Our framework combines nonparametric learning [20, 33] with the generalized moment method, which is mainly studied in the setting of parametric inference [76, 77, 85]. We study the identifiability of the observation function f_* from first-order moments, and show that the first-order generalized moments can identify the function in the L^2 closure of a system-intrinsic data-adaptive reproducing kernel Hilbert space (SIDA-RKHS). As far as we know, this is the first result on the function space of identifiability for nonparametric learning of observation functions in SSMs.

When the observation function is invertible, its unsupervised regression is investigated [78] by maximizing the likelihood for high-dimensional data. However, in many applications, particularly those involving complex dynamics, the observation functions are non-invertible, for example they are projections or nonlinear non-invertible transformations (e.g.,$f(x) = |x|^2$ with $x \in \mathbb{R}^d$). As a consequence, the resulting observed process may have discontinuous or singular probability densities [32, 42]. In [71], it has been shown empirically that delayed coordinates with principal component analysis may be used to estimate the dimension of the hidden process, and diffusion maps [18] may yield a diffeomorphic copy of the observation function.

The remainder of the Chapter is organized as follows. We present the nonparametric generalized moment method in Chapter 4.2. In Chapter 4.3 we study the identifiability of the observation function from first-order moments, and show that the function spaces of identifiability are RKHSs intrinsic to the state space model. We present numerical examples to demonstrate the effectiveness and the limitations of the proposed method in Chapter 4.4. Chapter 4.5 summarizes this study and discusses directions of future research.

2. Algorithm: nonparametric regression based on generalized moments

Throughout this chapter, we focus on discrete-time observations of the state space model (4.1) – (4.2), because in practice we observe data in discrete timestamps. We suppose that the data is in the form $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^{M}$, with m indexing multiple independent trajectories, observed at the vector $t_0 : t_L$ of discrete times $(t_0, \cdots, t_L)$. The extension to continuous time trajectories is straightforward.

2.1. Generalized moment method We estimate the observation function f_* by the generalized moment method (GMM), searching for an observation function $\widehat{f}$, in a suitable finite-dimensional hypothesis space, such that the moments of functionals of the process $\widehat{f}(X_t)$ are close to the empirical ones (computed from data) of $Y_t = f_*(X_t)$.

We consider "generalized moments" in the form $\mathbb{E}\left[\xi(Y_{t_0:t_L})\right]$, where $\xi : \mathbb{R}^{L+1} \to \mathbb{R}^K$ is a functional of the trajectory $Y_{t_0:t_L}$. The empirical generalized moments $E_M[\xi(Y_{t_0:t_L})]$ are computed from data by Monte Carlo approximation:

$$\mathbb{E}\left[\xi(Y_{t_0:t_L})\right] \approx E_M[\xi(Y_{t_0:t_L})] := \frac{1}{M}\sum_{m=1}^{M} \xi(Y_{t_0:t_L}^{(m)}), \tag{4.3}$$

which converges at the rate $M^{-1/2}$ by the Central Limit Theorem, since the M trajectories are independent. Meanwhile, since the distribution of the state process is known, for any putative observation function f, we approximate the moments of the process $(f(X_t)))$ by simulating M'

independent trajectories of the state process (X_t):

$$\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right] \approx \frac{1}{M'}\sum_{m=1}^{M'} \xi(f(X)_{t_0:t_L}^{(m)})\,. \tag{4.4}$$

Here, with some abuse of notation, $f(X)_{t_0:t_L}^{(m)} := (f(X_{t_0}^{(m)}), \ldots, f(X_{t_L}^{(m)}))$. The number M' can be as large as we can afford from a computational perspective. The calculations above can be done parallel over trajectories. Since M' can be chosen large – only subject to computational constraints – we consider the error in this empirical approximation negligible and work with $\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]$ directly.

We estimate the observation function f_* by minimizing a notion of discrepancy between these two empirical generalized moments:

$$\widehat{f} = \operatorname*{arg\,min}_{f\in\mathcal{H}} \mathcal{E}^M(f), \quad \text{where } \mathcal{E}^M(f) := \operatorname{dist}\left(E_M[\xi(Y_{t_0:t_L})], \mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]\right)^2, \tag{4.5}$$

where f is restricted to some suitable hypothesis space $\mathcal{H}$, and $\operatorname{dist}(\cdot,\cdot)$ is a suitable distance between the moments to be specified later. By the law of large numbers, $\mathcal{E}^M(f)$ tends almost surely to $\mathcal{E}(f) := \operatorname{dist}\left(\mathbb{E}\left[\xi(Y_{t_0:t_L})\right], \mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]\right)^2$.

It is desirable to choose the generalized moment functional ξ and the hypothesis space $\mathcal{H}$ so that the minimization in (4.5) can be performed efficiently. We choose $\mathcal{H}$ to be a subset of an n-dimensional function space, spanned by basis functions $\{\phi_i\}$, within which we can write $\widehat{f} = \sum_{i=1}^n \widehat{c}_i\phi_i$. We select the functional ξ so that the moments $\mathbb{E}\left[\xi(f(X)_{t_0:t_L})\right]$, for $f = \sum_{i=1}^n c_i\phi_i$, can be efficiently evaluated for all $(c_1, \ldots, c_n)$. To this end, we choose linear functionals or low-degree polynomials, so that we only need to compute the moments of the basis functions once, and use these moments repeatedly during the optimization process, as discussed in the next section.

The selection of the hypothesis space is detailed in Chapter 4 Section 2.3.

2.2. Loss functional and estimator The generalized moments we consider include the first and the second moments, and the one-step temporal correlation. We let

$$\xi(Y_{t_0:t_L}) := (Y_{t_0:t_L}, Y^2_{t_0:t_L}, Y_{t_0}Y_{t_1}, \ldots, Y_{t_{L-1}}Y_{t_L}) \in \mathbb{R}^{3L+2}.$$

The loss functional in (4.5) is then chosen in the following form: for weights $w_1, \ldots, w_3 > 0$,

$$\begin{aligned}\mathcal{E}(f) :=& w_1 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2}_{\mathcal{E}_1(f)} + w_2 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})^2] - \mathbb{E}[Y^2_{t_l}]\right|^2}_{\mathcal{E}_2(f)} \\ &+ w_3 \underbrace{\frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})f(X_{t_{l-1}})] - \mathbb{E}[Y_{t_l}Y_{t_{l-1}}]\right|^2}_{\mathcal{E}_3(f)}.\end{aligned} \tag{4.6}$$

In principle, these weights are selected to balance the contributions of these terms, and we set them according to data as detailed in (4.26).

Let the hypothesis space $\mathcal{H}$ be a subset of the span of a linearly independent set $\{\phi_i\}_{i=1}^n$, which we specify in the next section. For $f = \sum_{i=1}^n c_i\phi_i \in \mathcal{H}$, we can write the loss functionals $\mathcal{E}_1(f)$ in (4.6) as

$$\mathcal{E}_1(f) = \frac{1}{L}\sum_{l=1}^{L} \left| \sum_{i=1}^{n} c_i \mathbb{E}\left[\phi_i(X_{t_l})\right] - \mathbb{E}\left[Y_{t_l}\right] \right|^2 = c^\top \overline{A}_1 c - 2c^\top \overline{b}_1 + \tilde{b}_1, \tag{4.7}$$

where $\tilde{b}_1 := \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\left[Y_{t_l}\right]^2$, and the matrix $\overline{A}_1$ and the vector $\bar{b}_1$ are given by

$$\overline{A}_1(i,j) := \frac{1}{L}\sum_{l=1}^{L} \underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[\phi_j(X_{t_l})\right]}_{A_{1,l}(i,j)}, \qquad \bar{b}_1(i) := \frac{1}{L}\sum_{l=1}^{L} \underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[Y_{t_l}\right]}_{b_{1,l}(i)}. \tag{4.8}$$

Similarly, we can write $\mathcal{E}_2(f)$ and $\mathcal{E}_3(f)$ in (4.6) as

$$\begin{aligned}
\mathcal{E}_2(f) &= \frac{1}{L}\sum_{l=1}^{L}\left|\sum_{i=1}^{n} c_i c_j \underbrace{\mathbb{E}\left[\phi_i(X_{t_l})\phi_j(X_{t_l})\right]}_{A_{2,l}(i,j)} - \underbrace{\mathbb{E}\left[Y_{t_l}^2\right]}_{b_{2,l}}\right|^2, \\
\mathcal{E}_3(f) &= \frac{1}{L}\sum_{l=1}^{L}\left|\sum_{i=1}^{n} c_i c_j \underbrace{\mathbb{E}\left[\phi_i(X_{t_{l-1}})\phi_j(X_{t_l})\right]}_{A_{3,l}(i,j)} - \underbrace{\mathbb{E}\left[Y_{t_{l-1}}Y_{t_l}\right]}_{b_{3,l}}\right|^2.
\end{aligned} \tag{4.9}$$

Thus, with the above notations in (4.8)-(4.9), the minimizer of the loss functional $\mathcal{E}(f)$ over $\mathcal{H}$ is

$$\begin{aligned}
\widehat{f}_{\mathcal{H}} &:= \sum_{i=1}^{n} \widehat{c}_i \phi_i\,, \qquad \widehat{c} := \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^{n} c_i\phi_i \in \mathcal{H}}{\arg\min} \mathcal{E}(c), \qquad \text{where} \\
\mathcal{E}(c) &:= w_1[c^\top \overline{A}_1 c - 2c^\top \bar{b}_1 + \tilde{b}_1] + \sum_{k=2}^{3} w_k \frac{1}{L}\sum_{l=1}^{L}\left|c^\top A_{k,l} c - b_{k,l}\right|^2.
\end{aligned} \tag{4.10}$$

Here, with an abuse of notation, we denote $\mathcal{E}(\sum_{i=1}^{n} c_i\phi_i)$ by $\mathcal{E}(c)$.

In practice, we approximate the expectations involving the observation process (Y_t) by the corresponding empirical means with data $\{Y^{(m)}_{[t_1:t_N]}\}_{m=1}^{M}$, as in (4.3). Meanwhile, we approximate the expectations involving the state process (X_t) by Monte Carlo as in (4.4), using M' trajectories. We assume that the sampling errors in the expectations of (X_t), i.e. in the terms $\{A_{k,l}\}_{k=1}^{3}$, are negligible, since the basis $\{\phi_i\}$ can be chosen to be bounded functions (such as B-spline polynomials) and M' can be as large as we can afford. We approximate $\{b_{k,l}\}_{k=1}^{3}$ by their empirical means

$\{b_{k,l}^M\}_{k=1}^3$:

$$\begin{aligned} b_{1,l}(i) &= \mathbb{E}\left[\phi_i(X_{t_l})\right]\mathbb{E}\left[Y_{t_l}\right] &&\approx \mathbb{E}\left[\phi_i(X_{t_l})\right]\frac{1}{M}\sum_{m=1}^{M} Y_{t_l}^{(m)} &&=: b_{1,l}^M(i)\,, \\ b_{2,l} &= \mathbb{E}\left[|Y_{t_l}|^2\right] &&\approx \frac{1}{M}\sum_{m=1}^{M} |Y_{t_l}^{(m)}|^2 &&=: b_{2,l}^M\,, \\ b_{3,l} &= \mathbb{E}\left[Y_{t_{l-1}}Y_{t_l}\right] &&\approx \frac{1}{M}\sum_{m=1}^{M} Y_{t_{l-1}}^{(m)}Y_{t_l}^{(m)} &&=: b_{3,l}^M\,. \end{aligned} \tag{4.11}$$

The estimator from data is

$$\begin{aligned} \widehat{f}_{\mathcal{H},M} &= \sum_{i=1}^{n} \widehat{c}_i\phi_i, \qquad \widehat{c} = \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^n c_i\phi_i\in\mathcal{H}}{\arg\min} \mathcal{E}^M(c), \text{ where} \\ \mathcal{E}^M(c) &= w_1[c^\top \overline{A}_1 c - 2c^\top \overline{b}_1^M + \widetilde{b}_1^M] + \sum_{k=2}^{3} w_k \frac{1}{L}\sum_{l=1}^{L}\left|c^\top A_{k,l}c - b_{k,l}^M\right|^2, \end{aligned} \tag{4.12}$$

where $\overline{b}_1^M = \frac{1}{L}\sum_{l=1}^{L} b_{1,l}^M$ and $\widetilde{b}_1^M = \frac{1}{LM}\sum_{l=1}^{L}\sum_{m=1}^{M}\left(Y_{t_l}^{(m)}\right)^2$.

The minimization of $\mathcal{E}^M(c)$ can be performed with iterative algorithms efficiently since the data-based matrices and vectors, $\overline{A}_1, \overline{b}_1^M$ and $\{A_{k,l}, b_{k,l}^M\}_{k=2}^3$, only need to be computed once. The main source of sampling error is the empirical approximation of the moments of the process (Y_t). We specify the hypothesis space in the next section and provide a detailed algorithm for the computation of the estimator in Chapter 4 Section 2.4.

Remark 4.1 (Moments involving Itô's formula)**.** *When the data trajectories are continuous in time (or when they are sampled with a high frequency in time), we can utilize additional moments from Itô's formula. Recall that for $f \in C_b^2$, applying Itô formula for the diffusion process in* (4.1)*, we have*

$$f(X_{t+\Delta t}) - f(X_t) = \int_t^{t+\Delta t} \nabla f \cdot b(X_s)dW_s + \int_t^{t+\Delta t} \mathfrak{L}f(X_s)ds,$$

where the operator $\mathfrak{L}$ *is*

$$\mathfrak{L}f = \nabla f \cdot a + \frac{1}{2} Hess(f) : b^\top b. \tag{4.13}$$

Hence, $\mathbb{E}\left[\Delta Y_{t_l}\right] = \mathbb{E}\left[\mathfrak{L} f_*(X_{t_{l-1}})\right] \Delta t + o(\Delta t)$*, where* $\Delta Y_{t_l} = Y_{t_l} - Y_{t_{l-1}}$*. Thus, when* Δt *is small, we can consider matching the generalized moments*

$$\mathcal{E}_4(f) = \frac{1}{L} \sum_{l=1}^{L} \left| \mathbb{E}\left[\mathfrak{L} f(X_{t_{l-1}})\right] \Delta t - \mathbb{E}\left[\Delta Y_{t_l}\right] \right|^2. \tag{4.14}$$

Similarly, we can further consider the generalized moments $\mathbb{E}\left[Y_t \Delta Y_t\right]$ *and* $\mathrm{Var}(\Delta Y_t)$ *and the corresponding quartic loss functionals. Since they require the moments of the first- and second-order derivatives of the observation function, they are helpful when the observation function is smooth with bounded derivatives.*

2.3. Hypothesis space and optimal dimension We let the hypothesis space $\mathcal{H}$ be a class of bounded functions in $\mathrm{span}\{\phi_i\}_{i=1}^n$,

$$\mathcal{H} := \{f = \sum_{i=1}^{n} c_i \phi_i : y_{\min} \le f(x) \le y_{\max} \text{ for all } x \in \mathrm{supp}(\overline{\rho}_T)\}, \tag{4.15}$$

where the basis functions $\{\phi_i\}$ and density $\overline{\rho}_T$ are to be specified below, and the empirical bounds

$$y_{\min} := \min\{Y_{t_l}^{(m)}\}_{l,m=1}^{L,M}, \quad y_{\max} := \max\{Y_{t_l}^{(m)}\}_{l,m=1}^{L,M}$$

aim to approximate the upper and lower bounds for f_*. Here the dimension n will be selected adaptive to data to avoid under- and over-fitting, as detailed in Algorithm 4 in Appendix C.1. Note that the hypothesis space $\mathcal{H}$ is a bounded convex subset of the linear space $\mathrm{span}\{\phi_i\}_{i=1}^n$. While

the pointwise bound constraints are for all x, in practice, for efficient computation, we apply these constraints at representative points, for example at the mesh-grid points used when the basis functions are piecewise polynomials. One may apply stronger constraints, such as requiring time-dependent bounds to hold at all times: $y_{\min}(t) \le \sum_{i=1}^{n} c_i f_i(x) \le y_{\max}(t)$ for each time t, where $y_{\min}(t)$ and $y_{\max}(t)$ are the minimum and maximum of the data set $\{Y_t^{(m)}\}_{m=1}^{M}$.

Basis functions. We use B-spline basis $\{\phi_i\}$ consisting of piecewise polynomials for hypothesis space $\mathcal{H}$ (see Appendix A.2 for details). To specify the knots of B-spline functions, we introduce a density function $\overline{\rho}_T^L$, which is the average of the probability densities $\{p_{t_l}\}_{l=1}^{L}$ of $\{X_{t_l}\}_{l=1}^{L}$:

$$\overline{\rho}_T^L(x) = \frac{1}{L}\sum_{l=1}^{L} p_{t_l}(x) \quad \xrightarrow{L\to\infty} \overline{\rho}_T(x) = \frac{1}{T}\int_0^T p_t(x)dt, \tag{4.16}$$

when $t_L = T$ and $\max_{1\le l\le L} |t_l - t_{l-1}| \to 0$. Here $\overline{\rho}_T^L$ (and its continuous time limit $\overline{\rho}_T(x)$) describes the intensity of visits to the regions explored by the process (X_t). The knots of the B-spline function are from a uniform partition of $[R_{min}, R_{max}]$, the smallest interval enclosing the support of $\overline{\rho}_T^L$. Thus, the basis functions $\{\phi_i\}$ are piecewise polynomials with knots adaptive to the state space model which determines $\overline{\rho}_T^L$.

Dimension of the hypothesis space. It is important to select a suitable dimension of the hypothesis space to avoid under- or over-fitting. We select the dimension in two steps. First, we introduce an algorithm, *Cross-validating Estimation of Dimension Range* (CEDR), to estimate a proper range of the dimension from the quadratic loss functional $\mathcal{E}_1$. This avoids the sampling error amplification due to an unsuitably large dimension. The sampling error is estimated from data by splitting the data into two sets. Then, we select the optimal dimension that minimizes

the 2-Wasserstein distance between the measures of data and prediction. See Appendix C.2 for details. Here we use the 2-Wasserstein distance because it is sensitive to small changes in $\widehat{f}$ caused by overfitting and it can be efficiently computed for large-sample datasets.

2.4. Algorithm We summarize the above method of nonparametric regression with generalized moments in Algorithm 2. It minimizes a quartic loss function with the upper and lower bound constraints, and we perform the optimization with multiple initial conditions (see Appendix C.1 for details).

Input: The state space model and data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^{M}$ consisting of multiple trajectories of the observation process.
Output: Estimator $\widehat{f}$.
1: Estimate the empirical density $\overline{\rho}_T^L$ in (4.16) and find its support $[R_{min}, R_{max}]$.
2: Select a basis type, Fourier or B-spline, with an estimated dimension range $[1, N]$ (by Algorithm 4), and compute the basis functions using the support of $\overline{\rho}_T^L$, as described in Section 2.3, Chapter 4.
3: **for** $n = 1 : N$ **do**
4: Compute the moment matrices in (4.8)-(4.9) and the vectors $b_{k,l}^M$ in (4.11).
5: Find the estimator $\widehat{c}_n$ by optimization with multiple initial conditions. Compute and record the values of the loss functional and the 2-Wasserstein distances.
6: Select the optimal dimension n (and degree if B-spline basis) that has the minimal 2-Wasserstein distance in (C.5). Return the estimator $\widehat{f} = \sum_{i=1}^{n} c_n^i \phi_i$.

Algorithm 2: Estimating the observation function by nonparametric generalized moment methods

Computational complexity The computational complexity is driven by the construction of the normal matrix and vectors and the evaluation of the 2-Wasserstein distances, which have complexity of order $\mathrm{O}(n^2LM)$ and $\mathrm{O}(nLM)$, respectively, for an overall complexity $\mathrm{O}(n^2LM)$.

2.5. Tolerance to noise in the observations The generalized moment method can tolerate large additive observation noise if the distribution of the noise is known. The estimation error

caused by the noise is at the scale of the sampling error, which is negligible when the sample size is large.

More specifically, suppose that we observe $\{Y^{(m)}_{t_0:t_L}\}_{m=1}^M$ from the observation model

$$Y_{t_l} = f_*(X_{t_l}) + \eta_{t_l}, \tag{4.17}$$

where $\{\eta_{t_l}\}$ is sampled from a process (η_t) that is independent of (X_t) and has moments

$$\mathbb{E}[\eta_t] = 0, \quad C(s,t) = \mathbb{E}[\eta_t \eta_s], \text{ for } s,t \geq 0. \tag{4.18}$$

A typical example is when η being identically distributed independent Gaussian noise $\mathcal{N}(0,\sigma^2)$, which gives $C(s,t) = \sigma^2\delta(t-s)$.

The algorithm 2 can be applied to the noisy data with only a few small modifications. First, note that the loss functional in (4.6) involves only the moments $\mathbb{E}[Y_t]$, $\mathbb{E}[Y_t^2]$ and $\mathbb{E}[Y_{t_l}Y_{t_{l-1}}]$, which are moments of $f_*(X_t)$. When Y_t in (4.17) has observation noise specified above, we have

$$\begin{aligned}
\mathbb{E}[f_*(X_t)] &= \mathbb{E}[Y_t] - \mathbb{E}[\eta_t] = \mathbb{E}[Y_t]; \\
\mathbb{E}[f_*(X_t)f_*(X_s)] &= \mathbb{E}[Y_tY_s] - \mathbb{E}[\eta_t\eta_s] = \mathbb{E}[Y_tY_s] - C(t,s)
\end{aligned}$$

for all $t, s \geq 0$. Thus, we only need to change the loss functional to be

$$\begin{aligned}
\mathcal{E}(f) =& w_1 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2 + w_2 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})^2] - \mathbb{E}[Y_{t_l}^2] + C(t,t)\right|^2 \\
&+ w_3 \frac{1}{L}\sum_{l=1}^{L} \left|\mathbb{E}[f(X_{t_l})f(X_{t_{l-1}})] - \mathbb{E}[Y_{t_l}Y_{t_{l-1}}] + C(t,s)\right|^2.
\end{aligned} \tag{4.19}$$

Similar to (4.12), the minimizer of the loss functional can be then computed as

$$\widehat{f}_{\mathcal{H},M} = \sum_{i=1}^{n} \widehat{c}_i \phi_i, \quad \widehat{c} = \underset{c\in\mathbb{R}^n \text{ s.t. } \sum_{i=1}^n c_i\phi_i \in \mathcal{H}}{\arg\min} \mathcal{E}^M(c), \text{ where}$$

$$\begin{aligned} \mathcal{E}^M(c) &= w_1[c^\top \overline{A}_1 c - 2c^\top \overline{b}_1^M + \widetilde{b}_1^M] + w_2 \frac{1}{L}\sum_{l=1}^{L} \left| c^\top A_{2,l} c - b_{2,l}^M + C(t_l, t_l) \right|^2 \\ &+ w_3 \frac{1}{L}\sum_{l=1}^{L} \left| c^\top A_{3,l} c - b_{3,l}^M + C(t_l, t_{l-1}) \right|^2, \end{aligned} \tag{4.20}$$

where all the A-matrices and b-vectors are the same as before (in (4.8) – (4.9) and (4.11)).

Note that the observation noise introduces sampling errors through b_1^M, $b_{2,l}^M$ and $b_{2,l}^M$, which are at the scale $\mathrm{O}(\frac{1}{\sqrt{M}})$. Also, note the A-matrices are independent of the observation noise. Thus, the observation noise affects the estimator only through the sampling error at the scale $\mathrm{O}(\frac{1}{\sqrt{M}})$, the same as the sampling error in the estimator from noiseless data.

3. Identifiability analysis

In this section, we discuss the identifiability of the observation function by the loss functionals. We show that $\mathcal{E}_1$, the quadratic loss functional based on the first-order moments in (4.7), can identify the observation function in the $L^2(\overline{\rho}_T^L)$-closure of a reproducing kernel Hilbert space (RKHS) that is intrinsic to the state space model. In addition, the loss functional $\mathcal{E}_4$ in (4.14), based on the Itô formula, enlarges the function space of identifiability. We also discuss, in Chapter 4 Section 3.2, some limitations of the loss functional $\mathcal{E}$ in (4.19), which combines the quadratic and quartic loss functionals: in particular, symmetries or sampling from a stationary measure may prevent us from identifying the observation function when using only generalized moments.

We will specify such function spaces for $\mathcal{E}_1$, $\mathcal{E}_4$ and $\mathcal{E}_1 + \mathcal{E}_4$ in Chapter 4 Section 3.1. We note

that these function spaces do not take into account the constraints of upper and lower bounds, which generically lead to minimizers near or at the boundary of the constrained set. The loss functional $\mathcal{E}_2$ and $\mathcal{E}_3$ can be viewed as additional constraints.

3.1. Identifiability by quadratic loss functionals We consider the quadratic loss functionals $\mathcal{E}_1$ and $\mathcal{E}_4$, and show that they can identify the observation function in the $L^2(\overline{\rho}_T^L)$-closure of reproducing kernel Hilbert spaces (RKHSs) that are intrinsic to the state space model.

Assumption 4.2. *We make the following assumptions on the state space model.*

- *The coefficients in the state space model* (4.1) *satisfy global Lipschitz conditions, and therefore also a linear growth condition: there exists a constant* $C > 0$ *such that* $|a(x) - a(y)| + |b(x) - b(y)| \leq C|x - y|$ *for all* $x, y \in \mathbb{R}$, *and* $|a(x)| + |b(x)| \leq C(1 + |x|)$. *We assume that* $\inf_{x\in\mathbb{R}} b(x) > 0$. *Furthermore, we assume that* X_0 *has a bounded density.*

- *The observation function* f_* *satisfies* $\sup_{t\in[0,t_L]} \mathbb{E}\left[|f_*(X_t)|^2\right] < \infty$.

Theorem 4.3. *Given discrete-time data* $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^M$ *from the state space model* (4.1) *satisfying Assumption* 4.2, *let* $\mathcal{E}_1$ *and* $\mathcal{E}_4$ *be the loss functionals defined in* (4.6) *and* (4.14). *Denote* $p_t(x)$ *the density of the state process* X_t *at time* t, *and recall that* $\overline{\rho}_T^L$ *in* (4.16) *is the average, in time, of these densities. Let* $\mathfrak{L}^*$ *be the adjoint of the 2nd-order elliptic operator* $\mathfrak{L}$ *in* (4.13). *Then,*

(a) $\mathcal{E}_1$ *has a unique minimizer in* H_1, *the* $L^2(\overline{\rho}_T^L)$ *closure of the RKHS* $\mathcal{H}_{K_1}$ *with reproducing kernel*

$$K_1(x, x') = \frac{1}{\overline{\rho}_T^L(x)\overline{\rho}_T^L(x')} \frac{1}{L} \sum_{l=1}^{L} p_{t_l}(x) p_{t_l}(x'), \tag{4.21}$$

for (x, x') *such that* $\overline{\rho}_T^L(x)\overline{\rho}_T^L(x') > 0$, *and* $K_1(x, x') = 0$ *otherwise. When the data is continuous* ($L \to \infty$), *we have* $K_1(x, x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')} \frac{1}{T} \int_0^T p_t(x) p_t(x') dt$.

(b) *$\mathcal{E}_4$ has a unique minimizer in H_4, the $L^2(\overline{\rho}_T^L)$ closure of the RKHS $\mathcal{H}_{K_4}$ with reproducing kernel*

$$K_4(x,x') = \frac{1}{\overline{\rho}_T^L(x)\overline{\rho}_T^L(x')}\frac{1}{L}\sum_{l=1}^{L}\mathfrak{L}^* p_{t_l}(x)\mathfrak{L}^* p_{t_l}(x'), \tag{4.22}$$

for (x,x') such that $\overline{\rho}_T^L(x)\overline{\rho}_T^L(x') > 0$, and $K_4(x,x') = 0$ otherwise. When the data is continuous, we have $K_4(x,x') = \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T \mathfrak{L}^* p_t(x)\mathfrak{L}^* p_t(x')dt$.

(c) *$\mathcal{E}_1 + \mathcal{E}_4$ has a unique minimizer in H, the $L^2(\overline{\rho}_T^L)$ closure of the RKHS $\mathcal{H}_K$ with reproducing kernel*

$$K(x,x') = \frac{1}{\overline{\rho}_T^L(x)\overline{\rho}_T^L(x')}\frac{1}{L}\sum_{l=1}^{L}\left[p_{t_l}(x)p_{t_l}(x') + \mathfrak{L}^* p_{t_l}(x)\mathfrak{L}^* p_{t_l}(x')\right], \tag{4.23}$$

for (x,x') such that $\overline{\rho}_T^L(x)\overline{\rho}_T^L(x') > 0$, and $K(x,x') = 0$ otherwise. Similarly, we have $K(x,x') = \frac{1}{\bar{\rho}_T(x)\bar{\rho}_T(x')}\frac{1}{T}\int_0^T [p_t(x)p_t(x') + \mathfrak{L}^* p_t(x)\mathfrak{L}^* p_t(x')]dt$ for continuous data.

In particular, f_* is the unique minimizer of these loss functionals if f_* is in H_1, H_4 or H.

To prove this theorem, we first introduce an operator characterization of the RKHS $\mathcal{H}_{K_1}$ in the next lemma. Similar characterizations hold for the RKHSs $\mathcal{H}_{K_4}$ and $\mathcal{H}_K$.

Lemma 4.4. *The function K_1 in (4.21) is a Mercer kernel, that is, it is continuous, symmetric and positive semi-definite. Furthermore, K_1 is square integrable in $L^2(\overline{\rho}_T^L \times \overline{\rho}_T^L)$, and it defines a compact positive semi-definite integral operator $\mathcal{L}_{K_1} : L^2(\overline{\rho}_T^L) \to L^2(\overline{\rho}_T^L)$:*

$$\mathcal{L}_{K_1} h(x') = \int h(x)K_1(x,x')\overline{\rho}_T^L(x)dx. \tag{4.24}$$

Also, the RKHS $\mathcal{H}_{K_1}$ has the operator characterization: $\mathcal{H}_{K_1} = \mathcal{L}_{K_1}^{1/2}(L^2(\overline{\rho}_T^L))$ and $\{\sqrt{\lambda_i}\psi_i\}_{i=1}^{\infty}$ is an orthonormal basis of the RKHS $\mathcal{H}_{K_1}$, where $\{\lambda_i, \psi_i\}$ are the pairs of positive eigenvalues and corresponding eigenfunctions of $\mathcal{L}_{K_1}$.

Proof. Since the densities of diffusion process are smooth, the kernel K_1 is continuous on the support of $\overline{\rho}_T^L$ and it is symmetric. It is positive semi-definite (see Appendix A.1 for a definition) because for any $(c_1, \ldots, c_n) \in \mathbb{R}^n$ and $(x_1, \ldots, x_n)$, we have

$$\sum_{i,j=1}^{n} c_i c_j K(x_i, x_j) = \frac{1}{L} \sum_{l=1}^{L} \sum_{i,j=1}^{n} c_i c_j \frac{p_{t_l}(x_i) p_{t_l}(x_j)}{\overline{\rho}_T^L(x_i) \overline{\rho}_T^L(x_j)} = \frac{1}{L} \sum_{l=1}^{L} \left(\sum_{i=1}^{n} c_i \frac{p_{t_l}(x_i)}{\overline{\rho}_T^L(x_i)} \right)^2 \geq 0.$$

Thus, K_1 is a Mercer kernel.

To show that K_1 is square integrable, note first that $p_{t_l}(x) \leq \max_{1 \leq k \leq L} p_{t_k}(x) \leq L \overline{\rho}_T^L(x)$ for any x. Thus for each x, x', we have

$$\frac{1}{L} \sum_{l=1}^{L} p_{t_l}(x) p_{t_l}(x') \leq L^2 \overline{\rho}_T^L(x) \overline{\rho}_T^L(x')$$

and $K_1(x, x') \leq L$. It follows that K_1 is in $L^2(\overline{\rho}_T^L \times \overline{\rho}_T^L)$.

Since K_1 is positive definite and square integrable, the integral operator $\mathcal{L}_{K_1}$ is compact and positive semi-definite. The operator characterization follows from Theorem A.3 in Appendix A.1.

□

Remark 4.5. *The above lemma is only applicable to discrete-time observations because it uses the bound $K_1(x, x') \leq L$. When the data is continuous in time on $[0, T]$, we have $K_1 \in L^2(\overline{\rho}_T \times \overline{\rho}_T)$ if the support of $\overline{\rho}_T$ is compact. Since p_t is uniformly bounded above, $p_t(x) \leq \max_{y \in \mathbb{R}, s \in [0,T]} p_s(y) < \infty$, it is a regular solution of a Fokker-Planck equation which is uniformly elliptic by Assumption 4.2. Thus for each x, x', we have*

$$\frac{1}{T} \int_0^T p_t(x) p_t(x') dt \leq \left| \frac{1}{T} \int_0^T p_t(x)^2 \, dt \right|^{1/2} \left| \frac{1}{T} \int_0^T p_t(x')^2 \, dt \right|^{1/2} = \overline{\rho}_T(x)^{1/2} \overline{\rho}_T(x')^{1/2} \max_{y \in \mathbb{R}, s \in [0,T]} p_s(y)$$

by Cauchy-Schwartz for the first inequality. Then,

$$K_1(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T p_t(x)p_t(x')dt \le \overline{\rho}_T(x)^{-1/2}\overline{\rho}_T(x')^{-1/2}\max_{y\in\mathbb{R},s\in[0,T]} p_s(y).$$

It follows that K_1 is in $L^2(\overline{\rho}_T \times \overline{\rho}_T)$:

$$\int\int K_1^2(x,x')\overline{\rho}_T(x)\overline{\rho}_T(x')dxdx' \le |\mathrm{supp}(\overline{\rho}_T)|\max_{y\in\mathbb{R},s\in[0,T]} p_s(y)^2 < \infty.$$

When $\overline{\rho}_T$ has non-compact support, it remains to be proved that $K_1 \in L^2(\overline{\rho}_T \times \overline{\rho}_T)$.

Proof of Theorem 4.3. The proof for (a)–(c) are similar, so we focus on (a) and only sketch the proof for (b)–(c).

To prove (a), we only need to show the uniqueness of the minimizer, because Lemma 4.4 has shown that K_1 is a Mercer kernel. Furthermore, note that by Lemma 4.4, the $L^2(\overline{\rho}_T^L)$ closure of the RKHS $\mathcal{H}_{K_1}$ is $H_1 = \overline{\mathrm{span}\{\psi_i\}_{i=1}^\infty}$, the closure in $L^2(\overline{\rho}_T^L)$ of the eigenspace of $\mathcal{L}_{K_1}$ with non-zero eigenvalues, where $\mathcal{L}_{K_1}$ is the operator defined in (4.24).

For any $f \in H_1$, with the notation $h = f - f_*$, we have $\mathbb{E}[f(X_t)] - \mathbb{E}[Y_t] = \mathbb{E}[h(X_t)]$ for each t (recall that $Y_t = f_*(X_t)$). Hence, we can write the loss functional as

$$\begin{aligned}\mathcal{E}_1(f) =& \frac{1}{L}\sum_{l=1}^L \left|\mathbb{E}[f(X_{t_l})] - \mathbb{E}[Y_{t_l}]\right|^2 = \frac{1}{L}\sum_{l=1}^L \left|\mathbb{E}[h(X_{t_l})]\right|^2 \\ =& \int\int h(x)h(x')\frac{1}{L}\sum_{l=1}^L p_{t_l}(x)p_{t_l}(x')dxdx' \\ =& \int\int h(x)h(x')K_1(x,x')\overline{\rho}_T^L(x)\overline{\rho}_T^L(x')dxdx' \ge 0.\end{aligned} \tag{4.25}$$

Thus, $\mathcal{E}_1$ attains its unique minimizer in H_1 at f_* if and only if $\mathcal{E}_1(f_* + h) = 0$ with $h \in H_1$ implies

that $h = 0$. Note that the second equality in (4.25) implies that $\mathcal{E}_1(f_* + h) = 0$ if and only if $\mathbb{E}[h(X_{t_l})] = 0$, i.e. $\int h(x)p_{t_l}(x)dx = 0$, for all t_l. Then, $\int h(x)p_{t_l}(x)\frac{p_{t_l}(x')}{\overline{\rho}_T^L(x')}dx = 0$ for each t_l and x'. Thus, the sum of them is also zero:

$$0 = \int h(x)\frac{1}{L}\sum_{l=1}^{L}\frac{p_{t_l}(x)p_{t_l}(x')}{\overline{\rho}_T^L(x')\overline{\rho}_T^L(x)}\overline{\rho}_T^L(x)dx = \int h(x)K_1(x,x')\overline{\rho}_T^L(x)dx$$

for each x'. By the definition of the operator $\mathcal{L}_{K_1}$, this implies that $\mathcal{L}_{K_1}h = 0$. Thus, $h = 0$ because $h \in H_1$.

The above arguments hold true when the kernel K_1 is from continuous-time data: one only has to replace $\frac{1}{L}\sum_{l=1}^{L}$ by the averaged integral in time. This completes the proof for (a).

The proofs of (b) and (c) are the same as above except the appearance of the operator $\mathfrak{L}^*$. Note that $\mathcal{E}_4$ in (4.14) reads $\mathcal{E}_4(f) = \frac{1}{L}\sum_{l=1}^{L}\left|\mathbb{E}\left[\mathfrak{L}f(X_{t_l})\right] - \mathbb{E}\left[\Delta Y_{t_l}\right]\right|^2$, thus, it differs from $\mathcal{E}_1$ only at the expectation $\mathbb{E}\left[\mathfrak{L}f(X_{t_l})\right]$. By integration by parts, we have

$$\mathbb{E}\left[\mathfrak{L}f(X_s)\right] = \int \mathfrak{L}f(x)p_s(x)dx = \int f(x)\mathfrak{L}^*p_s(x)dx$$

for any $f \in C_b^2$. Then, the rest of the proof for Part (b) follows exactly as above with K_1 and $\mathcal{L}_{K_1}$ replaced by K_4 and $\mathcal{L}_{K_4}$. □

The following remarks highlight the implications of the above theorem. We consider only $\mathcal{E}_1$, but all the remarks apply also to $\mathcal{E}_4$ and $\mathcal{E}_1 + \mathcal{E}_4$.

Remark 4.6 (An operator view of identifiability)**.** *The unique minimizer of $\mathcal{E}_1$ in H_1 defined in Theorem 4.3 is the zero of its Fréchet derivative: $\widehat{f} = \mathcal{L}_{K_1}^{-1}\mathcal{L}_{K_1}f_*$, which is f_* if $f_* \in H_1$. In fact,*

we can write the loss functional $\mathcal{E}_1$ *as*

$$\mathcal{E}_1(f) = \langle f - f_*, \mathcal{L}_{K_1}(f - f_*)\rangle_{L^2(\overline{\rho}_T^L)}.$$

Thus, the Fréchet derivative of $\mathcal{E}_1$ *in* $L^2(\overline{\rho}_T^L)$ *is* $\nabla\mathcal{E}_1(f) = \mathcal{L}_{K_1}(f - f_*)$ *and we obtain the unique minimizer. Furthermore, this operator representation of the minimizer conveys two important messages about the inverse problem of finding the minimizer of* $\mathcal{E}_1$*: (1) it is ill-defined beyond* H_1*, and in particular, it is ill-defined on* $L^2(\overline{\rho}_T^L)$ *when* $\mathcal{L}_{K_1}$ *is not positive definite; (2) the inverse problem is ill-posed on* H_1*, because the operator* $\mathcal{L}_{K_1}$ *is compact and its inverse* $\mathcal{L}_{K_1}^{-1}$ *is unbounded.*

Remark 4.7 (Identifiability and normal matrix in regression)**.** *Suppose* $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ *and denote* $f = \sum_{i=1}^n c_i\phi_i$ *with* ϕ_i *being basis functions such as B-splines. As shown in* (4.7)-(4.8)*, the loss functional* $\mathcal{E}_1$ *is a quadratic function with normal matrix* $\overline{A}_1 = \frac{1}{L}\sum_{l=1}^L A_{1,l}$*,* $A_{1,l} = \mathbf{u}_l^\top \mathbf{u}_l$*, and* $\mathbf{u}_l = (\mathbb{E}[\phi_1(X_{t_l})], \ldots, \mathbb{E}[\phi_n(X_{t_l})]) \in \mathbb{R}^n$*. Thus, the rank of the matrix* $\overline{A}_1$ *is no larger than* $\min\{n, L\}$*. Note that* $\overline{A}_1$ *is the matrix approximation of* $\mathcal{L}_{K_1}$ *on the basis* $\{\phi_i\}_{i=1}^n$ *in the sense that*

$$\overline{A}_1(i,j) = \langle \mathcal{L}_{K_1}\phi_i, \phi_j\rangle_{L^2(\overline{\rho}_T^L)},$$

for each $1 \le i, j \le n$*. Thus, the minimum eigenvalue of* $\overline{A}_1$ *approximates the minimum eigenvalue of* $\mathcal{L}_{K_1}$ *restricted in* $\mathcal{H}_n$*. In particular, if* $\mathcal{H}_n$ *contains a nonzero element in the null space of* $\mathcal{L}_{K_1}$*, then the normal matrix will be singular; if* $\mathcal{H}_n$ *is a subspace of the* $L^2(\overline{\rho}_T^L)$ *closure of* $\mathcal{H}_{K_1}$*, then the normal matrix is invertible and we can find a unique minimizer.*

Remark 4.8 (Convergence of estimator)**.** *For a fixed hypothesis space, the estimator converges to the projection of* f_* *in* $\mathcal{H} \cap H_1$ *as the data size* M *increases, at the rate* $O(M^{-1/2})$*, with the error coming from the Monte Carlo estimation of the moments of observations. With data-adaptive*

hypothesis spaces, we are unable to prove the minimax rate of convergence as in classical nonparametric regression, due to the lack of a coercivity condition [56, 67]*, since the eigenvalues of the compact operator $\mathcal{L}_{K_1}$ converge to zero. A minimax rate would require an estimate on the spectral decay of $\mathcal{L}_{K_1}$, which we leave for future research.*

Remark 4.9 (Regularization using the RKHS)**.** *The RKHS H_{K_1} can be further utilized to provide a data-adaptive regularization norm in Tikhonov regularization (see Chapter 3.2).*

Examples of the RKHS. We emphasize that the reproducing kernel and the RKHS are intrinsic to the state space model (including the initial distribution). We demonstrate the kernels by analytically computing them when the process (X_t) is either the Brownian motion or the Ornstein-Uhlenbeck (OU) process. For simplicity, we consider continuous-time data. Recall that when the diffusion coefficient in the state space model (4.1) is a constant, the second-order elliptic operators $\mathfrak{L}$ is $\mathfrak{L}f = \nabla f \cdot a + \frac{1}{2}b^2\Delta f$, and its adjoint operator $\mathfrak{L}^*$ is

$$\mathfrak{L}^* p_s = -\nabla \cdot (a p_s) + \frac{1}{2} b^2 \Delta p_s,$$

where p_s denotes the probability density of X_s.

Example 4.10 (1D Brownian motion)**.** *Let $a = 0$ and $b = 1$. Assume $p_0(x) = \delta_{x_0}$, i.e., $X_0 = x_0$. Then, X_t is the Brownian motion starting from x_0 and $p_t(x) = \frac{1}{\sqrt{2\pi t}} e^{-\frac{(x-x_0)^2}{2t}}$ for each $t > 0$. We have $\overline{\rho}_T(x) = \frac{1}{T}\int_0^T p_t(x)dt = \frac{x-x_0}{T\sqrt{\pi}}\Gamma(-\frac{1}{2}, \frac{(x-x_0)^2}{2T})$ and*

$$K_1(x, x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')} \frac{1}{T} \int_0^T p_s(x) p_s(x') ds = \frac{T\Gamma(0, \frac{(x-x_0)^2+(x'-x_0)^2}{2T})}{2(x-x_0)(x'-x_0)\Gamma(-\frac{1}{2}, \frac{(x-x_0)^2}{2T})\Gamma(-\frac{1}{2}, \frac{(x'-x_0)^2}{2T})},$$

where $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}dt$ *is the upper incomplete Gamma function. Also, we have*

$$\mathfrak{L}^* p_s(x) = \phi_2(s,x)p_s(x), \text{ with } \phi_2(s,x) = \left(\frac{1}{s^2}(x-x_0)^2 - \frac{1}{s}\right).$$

Thus, the reproducing kernel K_4 *in* (4.22) *and* K *in* (4.23) *from continuous-time data are*

$$K_4(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T \phi_2(s,x)\phi_2(s,x')p_s(x)p_s(x')ds;$$
$$K(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T (1+\phi_2(s,x)\phi_2(s,x'))p_s(x)p_s(x')ds.$$

Example 4.11 (Ornstein-Uhlenbeck process)**.** *Let* $a(x) = \theta x$ *and* $b = 1$ *with* $\theta > 0$*. Assume* $p_0(x) = \delta_{x_0}$*, i.e.,* $X_0 = x_0$*. Then,* $X_t = e^{-\theta t}x_0 + \int_0^t e^{-\theta(t-s)}dB_s$*. It has a distribution* $\mathcal{N}(e^{-\theta t}x_0, \frac{1}{2\theta}(1-e^{-2\theta t}))$*, thus* $p_t(x) = \frac{1}{\sqrt{2\pi}\sigma_t}\exp(-\frac{(x-x_0^t)^2}{2\sigma_t^2})$*, where* $x_0^t := e^{-\theta t}x_0$ *and* $\sigma_t^2 := \frac{1}{2\theta}(1-e^{-2\theta t})$*. Computing the spatial derivatives, we have* $\mathfrak{L}^* p_s(x) = \frac{1}{2}\left[\frac{(x-x_0^s)^2}{\sigma_s^4} - \frac{1}{\sigma_s^2}\right]p_s(x) - (\theta x p_s(x))' = \phi_2(s,x)p_s(x)$*, where*

$$\phi_2(s,x) := \left[\frac{(x-x_0)^2}{2\sigma_s^4} - \frac{1}{2\sigma_s^2} - \theta + \frac{\theta}{\sigma_s^2}x(x-x_0^s)\right].$$

The reproducing kernels K_1 *in* (4.21)*,* K_4 *in* (4.22) *and* K *in* (4.23) *are*

$$K_1(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T p_s(x)p_s(x')ds;$$
$$K_4(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T \phi_2(s,x)\phi_2(s,x')p_s(x)p_s(x')ds;$$
$$K(x,x') = \frac{1}{\overline{\rho}_T(x)\overline{\rho}_T(x')}\frac{1}{T}\int_0^T (1+\phi_2(s,x)\phi_2(s,x'))p_s(x)p_s(x')ds.$$

In particular, when the process is stationary, we have $K_1(x,x') \equiv 1$ *and* $K_4(x,x') = 0$ *because*

$\mathfrak{L}^* p_s = 0$ *when* $p_s(x) = \frac{2\theta}{\sqrt{2\pi}} \exp(-\theta x^2)$ *is the stationary density.*

3.2. Non-identifiability due to stationarity and symmetry When the hypothesis space $\mathcal{H}$ has a dimension larger than the RKHS's, the quadratic loss functional $\mathcal{E}_1$ may have multiple minimizers. The constraints of upper and lower bounds, as well as the loss functionals $\mathcal{E}_2$ and $\mathcal{E}_3$, can help identifying the observation function. However, as we show next, identifiability may still not hold due to symmetry and/or stationarity.

Stationary processes. When the process (X_t) is stationary, we have limited information from the moments in our loss functionals. We have $\mathcal{E}_1(f) = |\mathbb{E}[Y_{t_1}] - \mathbb{E}[f(X_{t_1})]|^2$ with $K_1(x, x') \equiv 1$, so $\mathcal{E}_1$ can only identify a constant function. Also, the loss functional $\mathcal{E}_4$ is identically 0 because

$$\mathfrak{L}^* p_s = \partial_s p_s = 0 \quad \Leftrightarrow \quad \mathbb{E}[\mathfrak{L} h(X_s)] = 0 \text{ for any } h \in C_b^2.$$

In other words, the function space of identifiability with $\mathcal{E}_1 + \mathcal{E}_4$ is the space of constant functions. Meanwhile, the quartic loss functionals $\mathcal{E}_2$ and $\mathcal{E}_3$ also provide limited information: they become $\mathcal{E}_2 = \left|\mathbb{E}[f(X_{t_1})^2] - \mathbb{E}[Y_{t_1}^2]\right|^2$ and $\mathcal{E}_3 = |\mathbb{E}[f(X_{t_2}) f(X_{t_1})] - \mathbb{E}[Y_{t_2} Y_{t_1}]|^2$, the second-order moment and the temporal correlation at a single pair of times.

To see the ensuing limitations, consider the finite-dimensional hypothesis space $\mathcal{H}$ in (4.15). As in (4.12), with $f = \sum_{i=1}^n c_i \phi_i$, the loss functional becomes

$$\mathcal{E}(f) = c^\top \overline{A}_1 c - 2 c^\top \overline{b}_1^M + |\mathbb{E}[Y_{t_1}]|^2 + \sum_{k=2}^{3} \left| c^\top A_{k,1} c - b_{k,1}^M \right|^2,$$

where $\overline{A}_1$ is a rank-one matrix, and $\sum_{k=2}^3 |c^\top A_{k,1} c - b_{k,1}^M|^2$ only adds two additional constraints.

Thus, $\mathcal{E}$ has multiple minimizers in a linear space with dimension greater than 3. One has to resort to the upper and lower bounds in (4.15) for additional constraints, which lead to minimizers on the boundary of the resulting convex set.

Symmetry. When the distribution of the state process X_t is symmetric, a moment-based loss functional may not distinguish the true observation function from its symmetric counterpart. More specifically, if a transformation $r : \mathbb{R} \to \mathbb{R}$ preserves the distribution, i.e., $(X_t, t \geq 0)$ and $(r(X_t), t \geq 0)$ have the same distribution, then $\mathbb{E}[f(X_t)] = \mathbb{E}[f \circ r(X_t)]$ and $\mathbb{E}[f(X_t)f(X_s)] = \mathbb{E}[f \circ r(X_t) f \circ r(X_s)]$. Thus, our loss functional will not distinguish f from $f \circ r$. This is of course reasonable: the two functions yield the same observation process (in terms of the distribution), thus the observation data does not provide the information necessary for distinguishing f from $f \circ r$.

Example 4.12 (Brownian motion)**.** *Consider the standard Brownian motion X_t, whose distribution is symmetric about $x = 0$ (because the two processes $(X_t, t \geq 0)$ and $(-X_t, t \geq 0)$ have the same distribution). Let the transformation be $r(x) = -x$. Then, the two functions $f(x)$ and $f(-x)$ lead to the same observation process in distribution sense, thus they cannot be distinguished from the observations.*

4. Numerical results

We demonstrate the effectiveness and limitations of our algorithm using synthetic data in representative examples. The algorithm works well when the state space model's densities vary appreciably in time to yield a large function space of identifiability whose distance to the true observation function is small. In this case, our algorithm leads to a convergent estimator as the sample size

increases. We also demonstrate that when the state process (i.e., the Ornstein-Uhlenbeck process) is stationary or symmetric in distribution (i.e., the Brownian motion), the loss functional can have multiple minimizers in the hypothesis space, preventing us from identifying the observation functions (see Section 4.3, Chapter 4).

4.1. Numerical setup

Data generation. The synthetic data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^{M}$ with $t_l = l\Delta t$ are generated from the state space model, which is solved by the Euler-Maruyama scheme with a time-step $\Delta t = 0.01$ for $L = 100$ steps. We will consider sample sizes $M \in \{\lfloor 10^{3.5+j\Delta} \rfloor : j = 0, 1, 2, 3, 4,\ \Delta = 0.0625\}$ to test the convergence of the estimator.

To estimate the moments in the A-matrices and b-vectors in (4.8)–(4.9) by Monte Carlo, we generate a new set of independent trajectories $\{X_{t_l}^{(m)}\}_{m=1}^{M'}$ with $M' = 10^6$. We emphasize that these samples of X are independent of the data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^{M}$.

Inference algorithm. We follow Algorithm 3 to search for the global minimum of the loss functionals in (4.12). The weights for the $\mathcal{E}_k$'s are $w_k = L\sqrt{M}/\|m_k^Y\|$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^L$, and for $l = 0, 1, \cdots, L-1$,

$$m_k^Y(l) = \frac{1}{M}\sum_{m=1}^{M}(Y_{t_l}^{(m)})^k \text{ for } k = 1, 2 \quad \text{and} \quad m_3^Y(l) = \frac{1}{M}\sum_{m=1}^{M} Y_{t_l}^{(m)} Y_{t_{l+1}}^{(m)}\,. \tag{4.26}$$

For each example, we test hypothesis spaces, spanned by B-splines with degree in $\{0, 1, 2, 3\}$, with dimension in the range $[1, N]$, which is selected by Algorithm 4. We select the optimal dimension and degree with the minimal 2-Wasserstein distance between the predicted and true

distribution of Y. The details are presented in Appendix C.2.

Results assessment and presentation. We present three aspects of the estimator $\widehat{f}$:

- **Estimated and true functions.** We compare the estimator with the true function f_*, along with the $L^2(\overline{\rho}_T^L)$ projection of f_* to the linear space expanded by the elements of $\mathcal{H}$.
- **2-Wasserstein distance.** We present the 2-Wasserstein distance (see (C.5) for definition) between the distributions of $Y_{t_l} = f_*(X_{t_l})$ and $\widehat{f}(X_{t_l})$ for each time with training data and a new set of randomly generated data of size 10^6. The new (test) data has $Y_{t_l}^{(m)} = f_*(X_{t_l}^{(m)})$, i.e., the X's and Y's are generated in pairs, while in the training data the X's and Y's are generated independently. This pairing can lead to an effect on the 2-Wasserstein distance, which depends only on the empirical distribution of the samples, but such effect is negligible in our experiments due to the large sample size.
- **Convergence of $L^2(\overline{\rho}_T^L)$ error.** We test the convergence of the estimator in $L^2(\overline{\rho}_T^L)$ as the sample size M increases. The $L^2(\overline{\rho}_T^L)$ error is computed by the Riemann sum approximation. We present the mean and standard deviation of $L^2(\overline{\rho}_T^L)$ errors from 20 independent simulations. The convergence rate is also highlighted, and we compare it with the minimax convergence rate in classical nonparametric regression (see e.g., [33, 67]), which is $\frac{s}{2s+1}$ with $s-1$ being the degree of the B-spline basis. This minimax rate is not available yet for our method, see Remark 4.8.

4.2. Examples The state space model we consider is a stochastic differential equation with the double-well potential

$$dX_t = (X_t - X_t^3)dt + dB_t, X_{t_0} \sim p_{t_0} \tag{4.27}$$

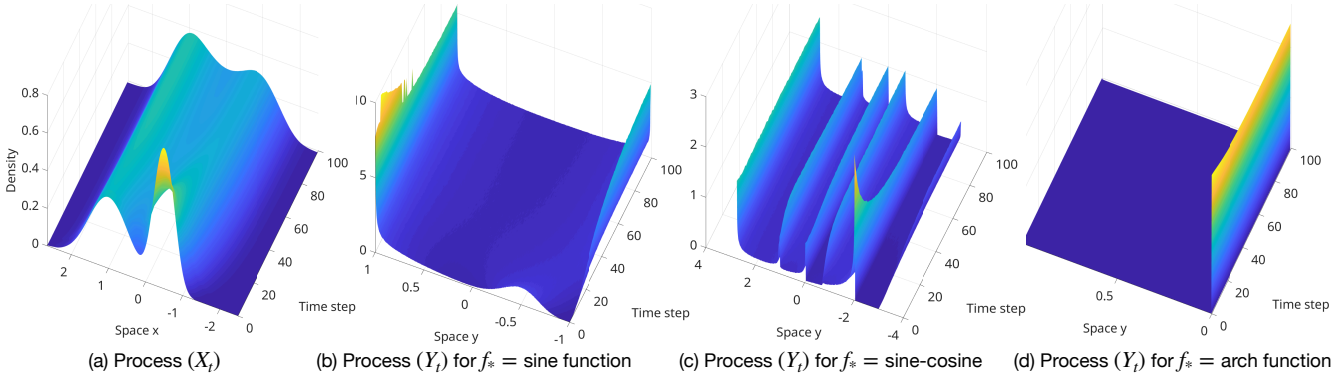
where the density of X_{t_0} is the average of $\mathcal{N}(-0.5, 0.2)$ and $\mathcal{N}(1, 0.5)$. The distribution of $X_{t_0:t_L}$ is non-symmetric and far from stationary (see Figure 4.1(a)); we therefore expect that the quadratic loss functional $\mathcal{E}_1$ provides a rich RKHS space for learning.

We consider three observation functions $f(x)$ representing typical challenges: nearly invertible, non-invertible, and non-invertible discontinuous, in $\text{supp}(\overline{\rho}_T)$:

$$\begin{aligned}
&\text{Sine function:} && f_1(x) = \sin(x); \\
&\text{Sine-Cosine function:} && f_2(x) = 2\sin(x) + \cos(6x); \\
&\text{Arch function:} && f_3(x) = \left(-2(1-x)^3 + 1.5(1-x) + 0.5\right) \mathbf{1}_{[0,1]}(x).
\end{aligned} \tag{4.28}$$

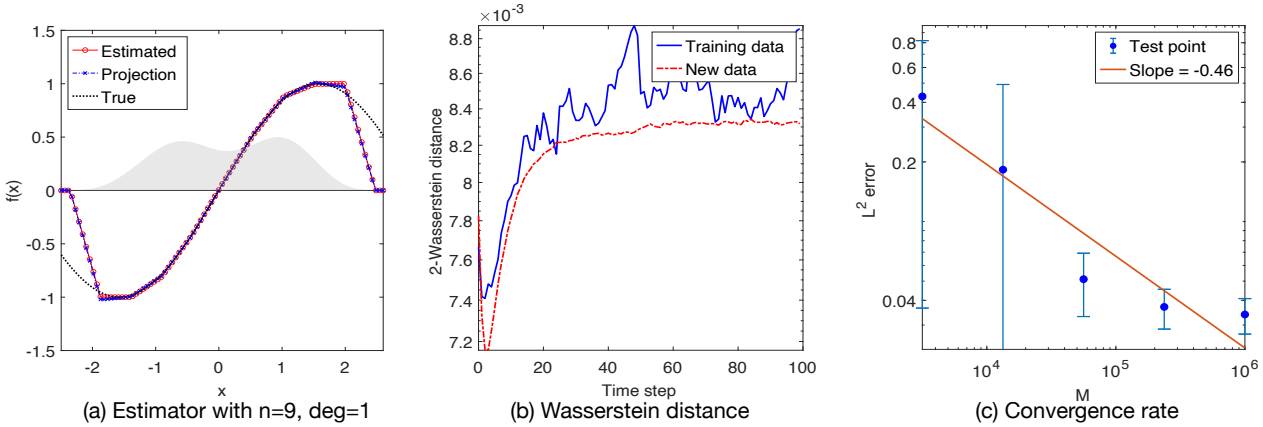
These functions are shown in Figure 4.2(a)–4.4(a). They lead to observation processes with dramatically different distributions, as shown in Figure 4.1(b-d).

The learning results for these three functions are shown in Figure 4.2–4.4. For each of these three observation functions, we present the estimator with the optimal hypothesis space, the 2-Wasserstein distance in prediction and the convergence of the estimator in $L^2(\overline{\rho}_T^L)$.



(a) Process (X_t) (b) Process (Y_t) for $f_* =$ sine function (c) Process (Y_t) for $f_* =$ sine-cosine (d) Process (Y_t) for $f_* =$ arch function

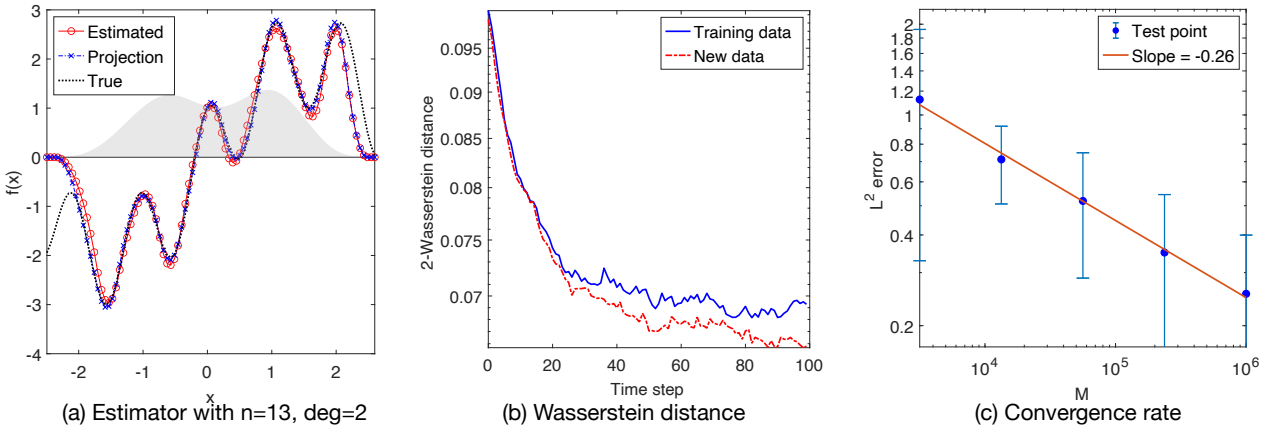
Figure 4.1: Empirical densities from the data trajectories of the state process (X_{t_l}) in double-well potential (4.27) and the observation processes $Y_{t_l} = f_i(X_{t_l})$.



(a) Estimator with n=9, deg=1 (b) Wasserstein distance (c) Convergence rate

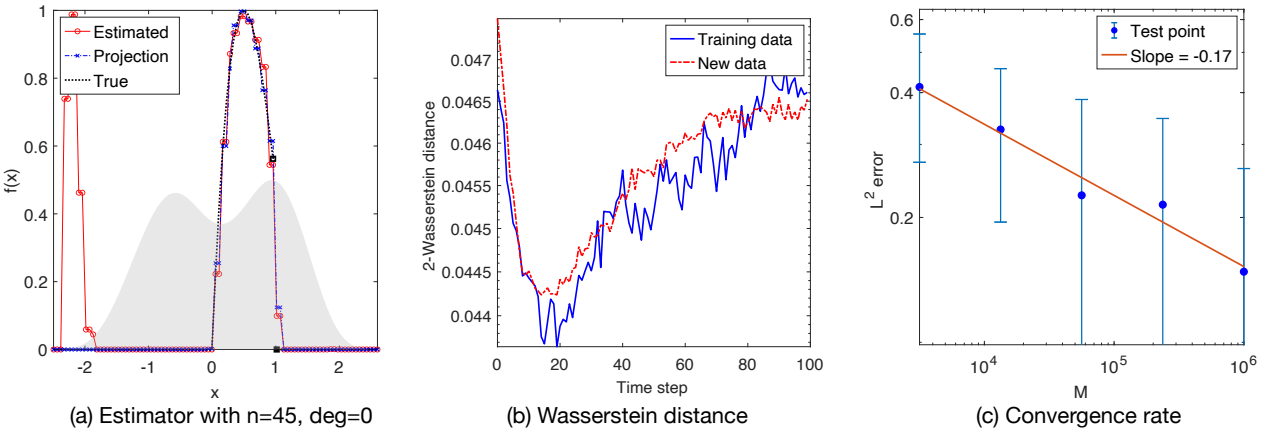
Figure 4.2: Learning results of Sine function f_1 with double-well potential (4.27).

Sine function: Figure 4.2(a) shows the estimator with degree-1 B-spline basis with dimension $n = 9$ for $M = 10^6$. The $L^2(\overline{\rho}_T^L)$ error is 0.0245 and the relative error is 3.47%. Figure 4.2(b) shows that the 2-Wasserstein distances are small at the scale 10^{-3}, in agreement with the sampling error since we used 10^6 samples. Figure 4.2(c) shows that the convergence rate of the $L^2(\overline{\rho}_T^L)$ error is 0.46. This rate is close to the minimax rate $\frac{2}{5} = 0.4$.



(a) Estimator with n=13, deg=2 (b) Wasserstein distance (c) Convergence rate

Figure 4.3: Learning results of Sine-Cosine function f_2 with double-well potential (4.27).



(a) Estimator with n=45, deg=0 (b) Wasserstein distance (c) Convergence rate

Figure 4.4: Learning results of Arch function f_3 with double-well potential (4.27).

Sine-Cosine function: Figure 4.3(a) shows the estimator with degree-2 B-spline basis with dimension $n = 13$. The $L^2(\overline{\rho}_T^L)$ error is 0.1596 and the relative error is 9.90%. The poor estimation near boundary is due to the lack of observation data. Figure 4.3(b) shows that the 2-Wasserstein distances are at the scale of 10^{-2}. Figure 4.3(c) shows that the convergence rate of the $L^2(\overline{\rho}_T^L)$ error is 0.26, less than the classical minimax rate $\frac{3}{7} \approx 0.42$. Note also that the variance of the L^2 error does not decrease as M increases. In comparison with the results for f_1 in Figure 4.2(a),

we attribute this relatively low convergence rate and the large variance to the high-frequency component $\cos(6x)$, which is harder to identify from moments than the low frequency component $\sin(x)$.

Arch function: Figure 4.4(a) shows the estimator with degree-0 B-spline basis with dimension $n = 45$. The $L^2(\overline{\rho}_T^L)$ error is 0.0645 and the relative error is 14.44%. Figure 4.4(b) shows that the 2-Wasserstein distances are small, at the scale 10^{-2}. Figure 4.4(c) shows that the convergence rate of the $L^2(\overline{\rho}_T^L)$ error is 0.17, less than the would-be minimax rate $\frac{1}{3} \approx 0.33$.

Arch function with observation noise: To demonstrate that our method can tolerate large observation noise, we present the estimation results from noisy observations of the Arch function, which is the most difficult one among the three examples. Suppose that the observation noise η in (4.17) is iid $\mathcal{N}(0, 0.25)$. Note that the average of $\mathbb{E}\left[|Y_t|^2\right]$ is about 0.2, so the signal-to-noise ratio is rather small, at $\mathbb{E}[|Y|^2]/\mathbb{E}[\eta^2] \approx 0.8$. Nevertheless, our method can identify the function using the moments of the noise as discussed in Chapter 4 Section 2.5. Figure 4.5(a) shows the estimator with degree-1 B-spline basis with dimension $n = 24$. The $L^2(\overline{\rho}_T^L)$ error is 0.1220 and the relative error is 27.32%. Figure 4.5(b) shows that the Wasserstein distances are small, of order 10^{-3}. The Wasserstein distances are approximated from samples of the noisy data $Y = f_{true}(X) + \eta$ and of the noisy prediction $\widehat{Y} = \widehat{f}(X) + \eta$. Figure 4.5(c) shows that the convergence rate of the $L^2(\overline{\rho}_T^L)$ error is 0.14. The estimation is not as good as the noise-free case. One reason would be that the noisy observation data lead to slightly lower and upper bound constraints in (4.15).

We consider this tolerance and robustness to noise to be quite surprising for such an ill-posed inverse problem, and the main reason for it is the use of moments methods, which average the noise so that the error occurs at scale $O(1/\sqrt{M})$.

We have also tested piecewise constant observation functions. Our method has difficulty in

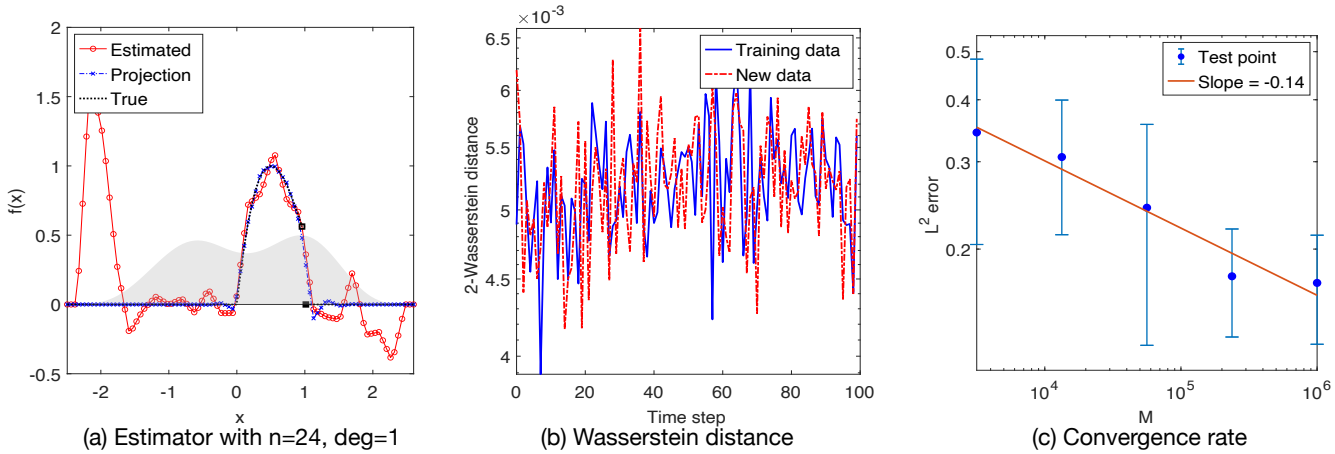


Figure 4.5: Learning results of Arch function f_3 with double-well potential (4.27) and i.i.d Gaussian observation noise.

identifying such functions, due to two issues: (i) the uniform partition often misses the jump discontinuities (even the projection of f_* has a large error); and (ii) the moments we considered depend on the observation function non-locally, thus, they provide limited information to identify the true function from its local perturbations. We leave it for future research to overcome these difficulties by searching the jump discontinuities and by introducing moments detecting local information.

4.3. Limitations We demonstrate by examples the non-identifiability due to symmetry and stationarity.

Symmetric distribution. Let the state space model be the Brownian motion with initial distribution $\mathrm{Unif}(0,1)$. The state process X_t has a distribution that is symmetric with respect to the line $x = \frac{1}{2}$. Thus, the processes X_t and $1 - X_t$ have the same distribution. Thus, with the reflection function $r(x) = 1 - x$, the processes $f(X_t)$ and $f \circ r(X_t)$ have the same distribution,

and the observation data does not provide information for distinguishing f from $f \circ r$. The loss functional (4.6) has at least two minima.

Figure 4.3 shows that our algorithm finds the reflection of the true function $f_* = \sin(x)$. The hypothesis space $\mathcal{H}$ has B-spline basis functions with degree 2 and dimension 58. Our estimator is close to $f_* \circ r(x) = \sin(1-x)$. Its $L^2(\overline{\rho}_T^L)$ error is 1.1244 and its reflection's $L^2(\overline{\rho}_T^L)$ error is 0.0790. Both the estimator and its reflection correctly predict the distribution of the observation process Y_t.

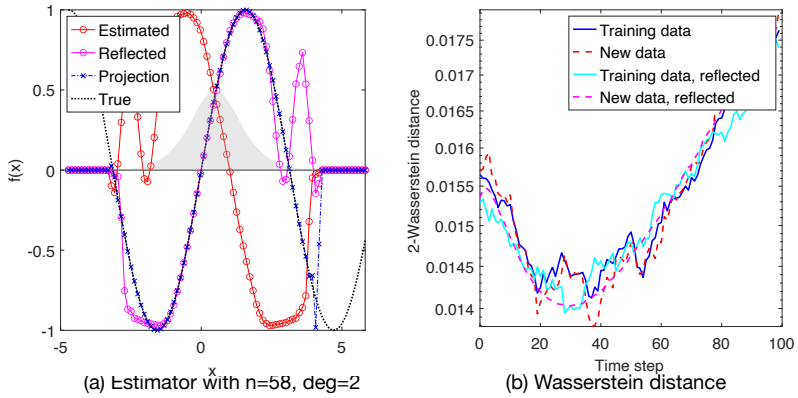


Figure 4.6: Learning results of $f_*(x) = \sin(x)$ with the SSM being $X_t = B_t + X_0$ where $X_0 \sim \text{Unif}(0,1)$. Due to the symmetry with respect to the line $x = \frac{1}{2}$, the estimator $\widehat{f}(x)$ and its reflection $\widehat{f}(1-x)$ are indistinguishable by the loss functional and they lead to similar prediction of the distribution of Y_{t_l}.

Stationary process. When the state process X_t is stationary, the loss functional (4.6) provides limited information about the observation function. As discussed in Section 3.2, the matrix $\overline{A}_1$ has rank 1, and $\mathcal{E}_2 = 0$ and $\mathcal{E}_3 = 0$ lead to only two more constraints. The constraints from the upper and lower bounds in (4.15) play a major role in leading to a minimizer at the boundary of the convex set $\mathcal{H}$.

Figure 4.3 shows the learning results with the stationary Ornstein-Uhlenbeck process $dX_t =$

$-X_t dt + dB_t$ and with the observation function $f_*(x) = \sin(x)$. The stationary density of (X_t) is $\mathcal{N}(0, \frac{1}{2})$. Due to the limited information, the estimator has a large $L^2(\overline{\rho}_T^L)$ error, which is 0.2656 and its prediction has large 2-Wasserstein distances oscillating near 0.1290.

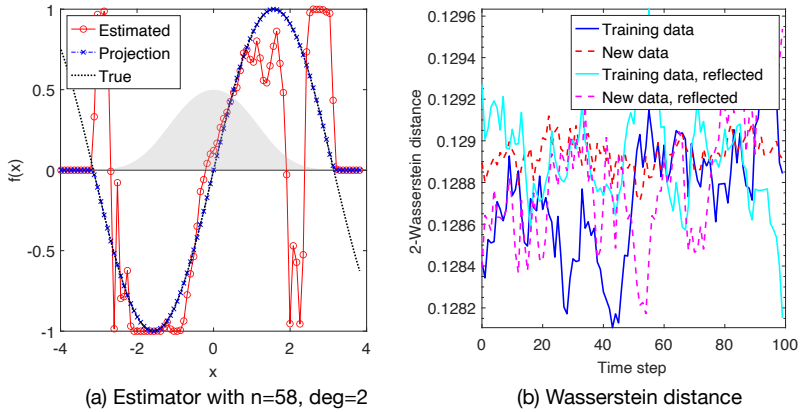


Figure 4.7: Learning results of $f_*(x) = \sin(x)$ with stationary Ornstein-Uhlenbeck process. Due to limited information from the moments, the estimator is inaccurate due to its reliance on the upper and lower bound constraints.

5. Discussions and conclusion

We have proposed a nonparametric learning method to estimate the observation functions in nonlinear state space models. It matches the generalized moments via constrained regression. The algorithm is suitable for large sets of unlabeled data. Moreover, it can deal with challenging cases when the observation function is non-invertible. We address the fundamental issue of identifiability from matching first-order moments. We show that the function spaces of identifiability are the closure of RKHS spaces intrinsic to the state space model. Numerical examples show that the first two moments and temporal correlations, along with upper and lower bounds, can identify smooth/non-smooth functions and tolerate considerable observation noise. The limitations of this

method, such as non-identifiability due to symmetry and stationarity, are also discussed.

This study provides a first step in the unsupervised learning of latent dynamics from abundant unlabeled data. There are several directions calling for further exploration: (i) a mixture of unsupervised and supervised learning that combines unlabeled data with limited labeled data, particularly for high-dimensional functions; (ii) enlarging the function space of learning, either by construction of more first-order generalized moments or by designing experiments to collect more informative data; (iii) joint estimation of the observation function and the state space model.

Appendix A

Preliminaries

1. A review of RKHS

We review the definitions and properties of positive semi-definite kernels. The following is a real-variable version of the definition in [8, p.67].

Definition A.1 (Positive semi-definite function)**.** *Let X be a nonempty set. A function $G : X \times X \to \mathbb{R}$ is positive semi-definite if and only if it is symmetric (i.e. $G(x, y) = G(y, x)$) and $\sum_{j,k=1}^{n} c_j c_k G(x_j, x_k) \geq 0$ for all $n \in \mathbb{N}$, $\{x_1, \ldots, x_n\} \subset X$ and $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}^n$. The function ϕ is positive definite if the equality hold only when $\mathbf{c} = \mathbf{0} \in \mathbb{R}^n$.*

Theorem A.2 (Properties of positive semi-definite kernels)**.** *Suppose that $k, k_1, k_2 : X \times X \subset \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ are positive semi-definite kernels. Then*

(a) $k_1 k_2$ is positive definite. ([8, p.69])

(b) The inner product $\langle u, v \rangle = \sum_{j=1}^{d} u_j v_j$ is positive semi-definite ([8, p.73])

(c) $f(u)f(v)$ is positive semi-definite for any function $f : X \to \mathbb{R}$ ([8, p.69]).

RKHS and positive integral operators. We review the definitions and properties of the Mercer kernel, the reproducing kernel Hilbert space (RKHS), and the related integral operator, see [20] for them on a compact domain [86] for them on a non-compact domain.

Let (X, d) be a metric space and $G : X \times X \to \mathbb{R}$ be continuous and symmetric. We say that G is a Mercer kernel if it is positive semi-definite. The RKHS H_G associated with G is defined to be closure of $\text{span}\{G(x, \cdot) : x \in X\}$ with the inner product

$$\langle f, g\rangle_{H_G} = \sum_{i=1,j=1}^{n,m} c_i d_j G(x_i, y_j)$$

for any $f = \sum_{i=1}^{n} c_i G(x_i, \cdot)$ and $g = \sum_{j=1}^{m} d_j G(y_j, \cdot)$. It is the unique Hilbert space such that: (1) the linear space $\text{span}\{G(\cdot, y), y \in X\}$ is dense in it; (2) it has the reproducing kernel property in the sense that for all $f \in H_G$ and $x \in X$, $f(x) = \langle G(x, \cdot), f\rangle_G$ (see [20, Theorem 2.9]).

By Mercer Theorem, we can characterize the RKHS H_G through the integral operator associated with the kernel. Let μ be a nondegenerate Borel measure on (X, d) (that is, $\mu(U) > 0$ for every open set $U \subset X$). Define the integral operator $\mathcal{L}_G$ on $L^2(X, \mu)$ by

$$\mathcal{L}_G f(x) = \int_X G(x, y) f(y) d\mu(y).$$

The RKHS has the operator characterization (see [20, Section 4.4] and [86]):

Theorem A.3. *Assume that the G is a Mercer kernel and $G \in L^2(X \times X, \mu \otimes \mu)$. Then*

1. *$\mathcal{L}_G$ is a compact positive self-adjoint operator. It has countably many positive eigenvalues $\{\lambda_i\}_{i=1}^{\infty}$ and corresponding orthonormal eigenfunctions $\{\phi_i\}_{i=1}^{\infty}$. Note that when zero is an eigenvalue of $\mathcal{L}_G$, the linear space $H = \text{span}\{\phi_i\}_{i=1}^{\infty}$ is a proper subspace of $L^2(\mu)$.*

2. *$\{\sqrt{\lambda_i}\phi_i\}_{i=1}^{\infty}$ is an orthonormal basis of the RKHS H_G.*

3. *The RKHS H_G is the image of the square root of the integral operator $\mathcal{L}_G$.*

2. B-spline basis

The choice of hypothesis space is important for the nonparametric regression. One can use global basis functions such as polynomials or Fourier basis when the function-valued parameter is known in prior to be smooth. On the other hand, when the function-valued parameter may be discontinuous, local basis functions such as B-splines or wavelets improve the estimation. In all our numerical experiments we choose the basis functions to be the B-splines, as we assume only limited information about the function-valued parameter.

B-Spline basis functions. We briefly review the definition of B-spline basis functions and we refer to [75, Chapter 2] and [70] for details. Given a nondecreasing sequence of real numbers, called knots, $(r_0, r_1, \ldots, r_m)$, the B-spline basis functions of degree p, denoted by $\{N_{i,p}\}_{i=0}^{m-p-1}$, are defined recursively as

$$N_{i,0}(r) = \begin{cases} 1, & r_i \le r < r_{i+1} \\ 0, & \text{otherwise} \end{cases}, \qquad N_{i,p}(r) = \frac{r - r_i}{r_{i+p} - r_i} N_{i,p-1}(r) + \frac{r_{i+p+1} - r}{r_{i+p+1} - r_{i+1}} N_{i+1,p-1}(r).$$

Each function $N_{i,p}$ is a nonnegative local polynomial of degree p, supported on $[r_i, r_{i+p+1}]$. At a knot with multiplicity k, it is $p - k$ times continuously differentiable. Hence, the differentiability increases with the degree but decreases when the knot multiplicity increases. The basis satisfies a partition unity property: for each $r \in [r_i, r_{i+1}]$, $\sum_j N_{j,p}(r) = \sum_{j=i-p}^{i} N_{j,p}(r) = 1$.

We set the knots of the spline functions to be a uniform partition of the support of the explo-

ration measure ρ. For any choice of degree p, we set the basis functions of the hypothesis space $\mathcal{H}$, contained in a subspace with dimension $n = m - p$, to be

$$\phi_i(r) = N_{i,p}(r),\ i = 0, \ldots, m - p - 1.$$

Thus, the basis functions $\{\phi_i\}$ are piecewise degree-p polynomials with knots adaptive to data.

Appendix B

1. Detailed nonparametric learning algorithm for learning kernels in operators

We consider only discrete data $\{u_k(x_j), v_k(x_j)\}_{k=1}^N$ in 1-dimensional and at equidistant mesh points $\{x_j = j\Delta x\}_{j=0}^J$. The extension to multi-dimensional and non-equidistant cases is straightforward.

Step 1: Estimate the exploration measure and assemble regression data.

We first estimate the exploration measure and extract the regression data that can be used for all hypothesis spaces by utilizing the regression structure and reading the data only once.

Let $d(\Omega)$ be the diameter of the set Ω. The discrete data set $\{u_k(x_j), v_k(x_j)\}_{k=1}^N$ explores only the variable r of ϕ in the set $\mathcal{R}_N^J = \{r_{ijk} = |y_i| \leq d(\Omega) : g[u_k](x_i, y_j) \neq 0 \text{ for some } i, j, k\}$, the set of all values explored by data with repetition. A discrete approximation of the exploration measure ρ in (3.2) is

$$\rho_N^J(dr) = \frac{1}{|\mathcal{R}_N^J|} \sum_{k=1}^{N} \sum_{i,j=1}^{J} \delta(|y_i| - r)|g[u_k](x_j, y_i)|. \tag{B.1}$$

This measure ρ_N^J uses only the information from u_k and it does not reflect the information about

the kernel in v_k.

We may estimate the support of the kernel from data. We set the data-adaptive support of the kernel to be $[0, R_0]$ with R_0 defined by

$$R_0 = 1.1 \min\{R_\rho, \max\{|L_i^v - L_i^u|, |R_i^v - R_i^u|\}_{i=1}^N\}, \tag{B.2}$$

where (L_i^u, R_i^u) and (L_i^v, R_i^v) are the lower and upper bounds of the supports $\text{supp}(u_k)$ and $\text{supp}(v_k)$ respectively, and R_ρ is the maximum of the support of ρ_N^J. That is, the support of the kernel lies inside the support of the exploration measure, and it is the maximal interaction range indicated by the difference between supports of u_k and v_k. Here the multiplicative factor 1.1 is an artificial factor to enlarge the range, so that the supports of the basis functions will fully cover the explored region.

Assemble regression data. Next, we assemble the regression data that will be used repeatedly, thus saving the computational cost by orders of magnitude, particularly when the data size is large with thousands of pairs (u_k, v_k). In order to compute the normal matrix $\overline{A}(i,j) = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle$ for any pair of basis functions, with the bilinear form defined in (3.3), we only need the integral kernel G. In particular, when $d = 1$, the integral $\int_{|\eta|=1} h(\eta) d\eta = h(\eta) + h(-\eta)$, therefore, we have

$$G(r,s) = \frac{1}{N} \sum_{k=1}^N \int_\Omega \left(g[u_k](x,r) + g[u_k](x,-r)\right) \left(g[u_k](x,s) + g[u_k](x,-s)\right) dx \tag{B.3}$$

for $r, s \in \text{supp}(\rho)$. Similarly, to compute $\overline{b}_n(i)$ in (3.11), which can be re-written as

$$\overline{b}_n(i) = \frac{1}{N} \sum_{k=1}^N \int R_{\phi_i}[u_k](x) v_k(x) dx = \int_0^R \phi_i(r) g_N^v(r) dr, \tag{B.4}$$

we only need the function g_N^v defined by

$$g_N^v(r) = \frac{1}{N}\sum_{k=1}^{N}\int_\Omega \left(g[u_k](x,r) + g[u_k](x,-r)\right) v_k(x)dx. \tag{B.5}$$

Let $r_k = k\Delta x$ for $k = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor$, which are the mesh points of ϕ explored by the data. Then, all the regression data we need in the original data (2.2) are

$$\left\{ G(r_k, r_l), g_N^v(r_k), \rho_N^J(r_k), \text{ with } k, l = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor \right\}, \tag{B.6}$$

where G, g_N^v and ρ_N^J are defined respectively in (B.3), (B.5) and (B.1).

Step 2: Select a class of hypothesis spaces and assemble regression matrices and vectors.

We set a class of data-adaptive hypothesis spaces $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ with their dimensions set to range from under-fitting to over-fitting. The basis functions can be either global basis functions such as polynomials and trigonometric functions, or local basis functions such B-spline polynomials (see [70] and Chapter 2 of [75]). To set the range for n, we note that the mesh points of the kernel's independent variable explored by data are $\{k\Delta x : k = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor\}$. Meanwhile, the basis function should be linearly independent in $L^2(\rho_N^J)$ so that the basis matrix

$$B_n = (\langle \phi_i, \phi_j \rangle_{L^2(\rho_N^J)})_{1 \le i,j \le n} \in \mathbb{R}^{n \times n} \tag{B.7}$$

is nonsingular. Thus, we set the range of n to be in $\lfloor \frac{R}{\Delta x} \rfloor \times [0.2, 1]$ such that B_n is nonsingular while covering a wide range of dimensions. For example, when we use piecewise constant basis, we can set $n = \lfloor \frac{R}{\Delta x} \rfloor$ with $\phi_i(x) = \delta(x_i - x)$, and we get $B_n = \text{Diag}(\rho_N^J)$. Thus, we estimate

the kernel as a vector of its values on the mesh points, with $L^2(\rho_N^J)$ being a vector space with a discrete-measure ρ_N^J.

With these regression data, the triplet $(\overline{A}_n, \overline{b}_n, B_n)$ can be efficiently evaluated for any basis functions using a numerical integrator to approximate the corresponding integrals. For example, with Riemann sum approximation, we compute the normal matrix $\overline{A}_n$ and vector $\overline{b}_n$ and the basis matrix B_n as

$$\begin{aligned}
\overline{A}_n(i,j) &= \langle\!\langle \phi_i, \phi_j \rangle\!\rangle \approx \sum_{k,l} \phi_i(r_k)\phi_j(r_l)G(r_k, r_l))\Delta x^2, \\
\overline{b}_n(i) &\approx \sum_k \phi_i(r_k) g_N^v(r_k))\Delta x, \\
B_n(i,j) &\approx \sum_k \phi_i(r_k)\phi_i(r_k)\phi_j(r_k)\rho_N^J(r_k)\Delta x.
\end{aligned} \tag{B.8}$$

The triplet $(\overline{A}_n, \overline{b}_n, B_n)$ is all we need for regression with regularization in the next step.

Step 3: Regression with DARTR.

Our DARTR method uses the norm of the SIDA-RKHS, which is the function space of identifiability as discussed in Chapter 3.1. That is, our estimator is the minimizer of the regularized loss in (1.10) with the regularization norm $\mathcal{R}(\phi) = \|\phi\|_{H_G}^2$ defined in (3.8).

Computation of the RKHS norm In practice, we can effectively approximate the RKHS norm $\|\phi\|_{H_G}^2$ using the triplet $(\overline{A}_n, \overline{b}_n, B_n)$. It proceeds in three steps. First, we solve the generalized eigenvalue problem $\overline{A}_n Q = B_n Q \Lambda$, where Λ is a diagonal matrix of the generalized eigenvalues and the matrix Q has columns being eigenvectors orthonormal in the sense that $Q^\top B_n Q = I_n$. Here these eigenvalues approximate the eigenvalue of $\mathcal{L}_{\overline{G}}$ in (3.6), and $\widehat{\psi}_k = \sum_j Q_{jk}\phi_j$ approximates the eigenfunctions of $\mathcal{L}_{\overline{G}}$. Then, we compute the square RKHS norm of $\phi = \sum_i c_i \phi_i$ as

$$\|\phi\|_{H_G}^2 = c^\top B_{rkhs} c, \text{ with } B_{rkhs} = (Q\Lambda Q^\top)^{-1}, \tag{B.9}$$

where the inverse is taken as pseudo-inverse, particularly when Λ has zero eigenvalues.

With the RKHS-norm ready, we write the regularized loss for each function $\phi = \sum_i c_i \phi_i$ as

$$\mathcal{E}_\lambda(\phi) = c^\top (\overline{A}_n + \lambda B_{rkhs}) c - 2c^\top \overline{b}_n + C_N^v.$$

The regularized estimator is

$$\widehat{\phi_\lambda} = \sum_{i=1}^{n} c_\lambda^i \phi_i, \;\; c_\lambda = (\overline{A}_n + \lambda B_{rkhs})^{-1} \overline{b}_n. \tag{B.10}$$

Then, we select the hyper-parameter λ by the L-curve method (see Appendix B.2).

Remark B.1 (Least squares to avoid matrix inverse)**.** *The matrix inverses can cause numerical issues when the normal matrix $\overline{A}$ is ill-conditioned or singular. Fortunately, the matrix inversions in B_{rkhs} and in solving $(\overline{A}_n + \lambda B_{rkhs}) c_\lambda = \overline{b}_n$ can be avoided by using minimum norm least squares solution. Note that this linear equation is equivalent to $(B_{rkhs}^{-T/2} \overline{A}_n B_{rkhs}^{-1/2} + \lambda I) \widetilde{c}_\lambda = B_{rkhs}^{-T/2} \overline{b}_n$ with $\widetilde{c}_\lambda = B_{rkhs}^{-1/2} c_\lambda$, where $B_{rkhs}^{-T/2}$ is the transpose of the square root matrix $B_{rkhs}^{-1/2}$. Meanwhile, the square root $B_{rkhs}^{-1/2} = (Q \Lambda Q^\top)^{1/2}$ comes directly from* (B.9)*. Thus, these treatments avoid the matrix inversions and lead to more robust estimators.*

We summarize the method in Algorithm 3.

Input: The data $\{u_k, v_k\}_{k=1}^N = \{u_k(x_j), v_k(x_j)\}_{k,j=1}^{N,J}$ with $x_j = j\Delta x$ to construct the nonlocal model $R_\phi[u] = f$.

Output: Estimator $\widehat{\phi}$

1: Estimate the exploration measure ρ_N^J from data as in (B.1), and estimate the support of the kernel from data as in (B.2). Let R be the upper bound of the support.

2: Get regression data (G, g_N^v) in (B.6).

3: Select a class of hypothesis spaces $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ by selecting a type of basis functions, e.g., polynomials or B-splines, n in the range $\lfloor \frac{R}{\Delta x} \rfloor \times [0.2, 1]$.

4: For each n, compute $(\overline{A}_n, \overline{b}_n, B_n)$ as in (B.8) for $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$, using (G, g_N^v, ρ_N^J) obtained above. If the basis matrix B_n is singular, remove n from the range. For the $(\overline{A}_n, \overline{b}_n, B_n)$, find the best regularized estimator $\widehat{c}_{\lambda_n}$ by DARTR in Algorithm 1, as well as corresponding loss value $\mathcal{E}(\widehat{c}_{\lambda_n})$.

5: Select the optimal dimension n^* (and degree if using B-spline basis) that has the minimal loss value (along with other cross-validation criteria if available). Return the estimator $\widehat{\phi} = \sum_{i=1}^{n^*} c_{n^*}^i \phi_i$.

Algorithm 3: Nonparametric learning of the nonlocal kernel with spare-aware regularization

2. Hyper-parameter by the L-curve method

We select the parameter λ by the L-curve method [36, 53]. Let l be a parametrized curve in $\mathbb{R}^2$:

$$l(\lambda) = (x(\lambda), y(\lambda)) := (\log(\mathcal{E}(\widehat{\phi_\lambda}), \log(\mathcal{R}(\widehat{\phi_\lambda})),$$

where $\mathcal{E}(\widehat{\phi_\lambda}) = c_\lambda^\top \overline{A}_n c_\lambda - 2c_\lambda^\top \overline{b}_n - C_N^v$, and $\mathcal{R}(\phi)$ is the regularization term, for example, $\mathcal{R}(\widehat{\phi_\lambda}) = \|\widehat{\phi_\lambda}\|_{H_G}^2 = c_\lambda^\top B_{rkhs} c_\lambda$. The optimal parameter is the maximizer of the curvature of l. In practice,

we restrict λ in the spectral range $[\lambda_{min}, \lambda_{max}]$ of the operator $\mathcal{L}_{\overline{G}}$,

$$\lambda_0 = \underset{\lambda_{\min} \le \lambda \le \lambda_{\max}}{\arg\max} \ \kappa(l(\lambda)) = \underset{\lambda_{\min} \le \lambda \le \lambda_{\max}}{\arg\max} \ \frac{x'y'' - x'y''}{(x'^2 + y'^2)^{3/2}}, \tag{B.11}$$

where λ_{min} and λ_{max} are computed from the smallest and the largest generalized eigenvalues of $(\overline{A}_n, B_n)$. This optimal parameter λ_0 balances the loss $\mathcal{E}$ and the regularization (see [36] for more details). In practice, instead of computing the second order derivatives, we compute the curvature by the reciprocal of the radius of the interior circle of three consecutive points[1].

We note that the performance of these regularizers depends on the optimal regularization strength λ_0. In our tests, all regularizers can successfully select the optimal λ_0 for most of the time, and the SIDA-RKHS regularizer has the most well-shaped L-curve, which leads to the most robust regularization (see Figure B.2 for typical L-curve plots).

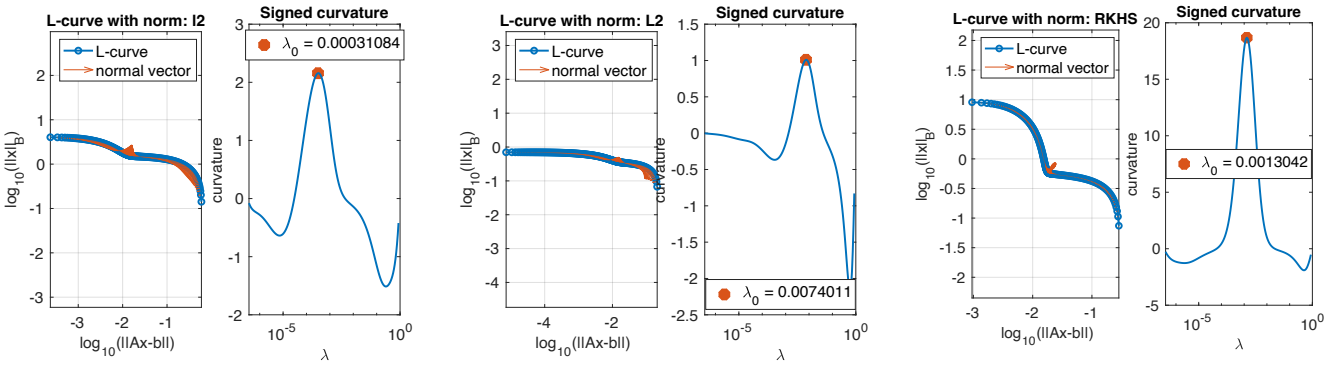


Figure B.1: Typical L-curve plots for the selection of the optimal regularization parameter λ_0 for the Gaussian kernel with $\Delta x = 0.05$ and nsr $= 1$.
From left to right: the l^2, L^2 and SIDA-RKHS regularizers. All regularizers successfully select the optimal λ_0, and the SIDA-RKHS regularizer has the most well-shaped L-curve.

[1] Are Mjaavatten (2022). Curvature of a 1D curve in a 2D or 3D space (`https://www.mathworks.com/matlabcentral/fileexchange/69452-curvature-of-a-1d-curve-in-a-2d-or-3d-space`), MATLAB Central File Exchange.

3. Detailed real-world dataset experiment settings

In this section we provide further experiment details for the real-world dataset studied in 3.4.

For both training and validation purposes we generate data using high-fidelity (HF) simulations for the propagation of stress waves within the microstructure of the heterogeneous, linear elastic bar. In the following, we use $\hat{u}$ to denote the HF solution, to distinguish the HF dataset from the homogenized solution of (3.16). The HF-model is a classical wave equation: the displacement $\hat{u}(x,t)$ satisfies, for $(x,t) \in \Omega \times [0,T]$ with $\Omega \subset \mathbb{R}$,

$$\partial_{tt}\hat{u}(x,t) - R_{HF}[\hat{u}](x,t) = g(x,t), \tag{B.12}$$

with a force loading term $g(x,t)$, proper boundary conditions and initial conditions $\hat{u}(x,0) = 0$, $\partial_t \hat{u}(x,0) = 0$. Considering the heterogeneous bar of two materials depicted in Figure 3.4, (B.12) describes the stress wave propagating with speed $c_1 = \sqrt{E_1/\rho}$ in material 1 and speed $c_2 = \sqrt{E_2/\rho}$ in material 2. We solve the HF-model (B.12) by the direct numerical solver (DNS) introduced in [83]. The DNS employs the characteristic line method, which provides exact solutions of velocities. For each grid point $x_j \in \Omega$ at time step $t^n = n\Delta t$, where Δt is the time step size, with the calculated exact velocity $\hat{v}(x_j, t^n)$ and the estimated displacement from the last time step $\hat{u}(x_j, t^{n-1})$ we update the HF displacement by

$$\hat{u}(x_j, t^n) = \hat{u}(x_j, t^{n-1}) + \Delta t \hat{v}(x_j, t^n).$$

With the above procedure, we then consider various boundary velocity loading $\partial_t \hat{u}_i(x,t)$, $x \in \partial\Omega$, and force loading $g_i(x,t)$ scenarios, and solve for the corresponding HF displacement field $\hat{u}_i(x,t)$. Resultant data pairs $\{\hat{u}_i, g_i\}_{i=1}^N = \{\hat{u}_i(x_j,t^n), g_i(x_j,t^n) : j = 1,\dots,J\}_{i=1,n=0}^{N,T/\Delta t}$ are employed as the

training and validation datasets. Discretization parameters for the DNS solver are set to $\Delta t = 0.01$ and $\max \Delta x = 0.01$.

The homogenization problem is then to learn the kernel of the nonlocal operator R_ϕ that approximates the operator R_{HF} from data $\{\hat{u}, v\}$ generated by $R_{HF}[\hat{u}] = v$, where $v = \partial_{tt}\hat{u} - g$. Discretizing the time derivative in (3.16) with the central difference scheme, we obtain

$$\frac{1}{\Delta t^2}(\hat{u}^{n+1}(x) - 2\hat{u}^n(x) + \hat{u}^{n-1}(x)) - g(x, t^n) := v^n(x),$$

where $\hat{u}^n(\cdot) := \hat{u}(\cdot, t^n)$ denotes the solution at time t^n. Given $\mathcal{D} = \{\hat{u}_i^n(x), v_i^n(x)\}_{i=1,n=1}^{N,T/\Delta t}$, our goal is to learn the kernel ϕ. The loss functional is

$$\mathcal{E}(\phi) = \frac{\Delta t}{NT} \sum_{k=1}^{N} \sum_{n=1}^{T/\Delta t} \|R_\phi[\hat{u}_k^n] - v_k^n\|_{L^2(\Omega)}^2. \tag{B.13}$$

3.1. Settings on real-world data In the learning problem, we consider four types of data and use the first three for training and the last one for validation of our algorithm. For all data we set $L = 0.2$, $\Delta t = 0.02$, $E_1 = 1$, $E_2 = 0.25$, $\rho = 1$, and the symmetric domain $\Omega = [-b, b]$. The estimated support of the kernel has a bound $R = 1.65$. Two spatial resolutions, $\Delta x = 0.05$ and $\Delta x = 0.025$ are considered, which we denote as the "coarse" and "fine" datasets, respectively.

Type 1 *Oscillating source (20 samples in total).*

$b = 50$, $T = 2$, $g(x,t) = \exp^{-(\frac{2x}{5jL})^2} \exp^{-(\frac{t-0.8}{0.8})^2} \cos^2(\frac{2\pi x}{jL})$, where $j = 1, 2, \cdots, 20$.

Type 2 *Plane wave with* cos *loading (11 samples in total).*

$b = 50$, $T = 2$, $g(x,t) = 0$ and $\partial_t u(-50, t) = \cos(jt)$, where the loading frequency $j = 0.35, 0.70, \cdots, 3.85$.

Type 3 *Plane wave with* sin *loading (11 samples in total).* $b = 50$, $T = 2$, $g(x,t) = 0$ and $\partial_t u(-50,t) = \sin(jt)$, where the loading frequency $j = 0.35, 0.70, \cdots, 3.85$.

Type 4 *Wave packet (3 samples in total).* $b = 133.3$, $T = 100$, $g(x,t) = 0$ and $\partial_t u(-b,t) = \sin(jt)\exp\left(-(t/5-3)^2\right)$, for $j = 1,\ 2,\ 3$.

Notice that the validation dataset (Type 4 dataset) is under a different loading condition from the training dataset, and with a much longer simulation time.

Appendix C

1. Detailed nonparametric learning algorithm for learning observation functions in SSMs

1.1. Dimension of the hypothesis space. The choice of dimension n of hypothesis space is important to avoid under- and over-fitting. We choose it by minimizing the 2-Wasserstein distance between the empirical distributions of observed process (Y_t) and that predicted by our estimated observation function. We proceed in 2 steps: first we determine a rough range for n, and then within this range we select the dimension with the minimal Wasserstein distance.

Step 1: We introduce an algorithm, called *Cross-validating Estimation of Dimension Range (CEDR)*, to estimate the range $[1, N]$ for the dimension of the hypothesis space, based on the quadratic loss functional $\mathcal{E}_1$. Its main idea is to restrict N to avoid overly amplifying the estimator's sampling error, which is estimated by splitting the data into two sets. It incorporates the function space of identifiability in Chapter 4 Section 3.1 into the SVD analysis [26, 36] of the normal matrix and vector from $\mathcal{E}_1$.

The CEDR algorithm estimates the sampling error in the minimizer of loss functional $\mathcal{E}_1$

through SVD analysis in three steps. First, we compute the normal matrix $\overline{A}_1$ and vector $\overline{b}_1$ in (4.8) by Monte Carlo; to estimate the sampling error in $\overline{b}_1$, we compute two copies, b and b', of $\overline{b}_1$ from two halves of the data:

$$b(i) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\left[\phi_i(X_{t_l})\right] \frac{2}{M} \sum_{m=1}^{\lfloor \frac{M}{2} \rfloor} Y_{t_l}^{(m)}, \quad b'(i) = \frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\left[\phi_i(X_{t_l})\right] \frac{2}{M} \sum_{m=\lfloor \frac{M}{2} \rfloor+1}^{M} Y_{t_l}^{(m)}. \tag{C.1}$$

Second, we implement an eigen-decomposition to find an orthonormal basis of $L^2(\overline{\rho}_T^L)$, the default function space of learning. The matrix $\overline{A}_1$ is a representation of the integral operator $\mathcal{L}_{K_1}$ in Lemma 3.2 on $\mathcal{H} = \text{span}\{\phi_i\}_{i=1}^n$, and $\mathcal{L}_{K_1}$'s eigenvalues are solved by the generalized eigenvalue problem

$$\overline{A}_1 u = \lambda B u, \qquad \text{where } B = (\langle \phi_i, \phi_j \rangle_{L^2(\overline{\rho}_T^L)}) \tag{C.2}$$

(see [52, Theorem 5.1]). Denote the eigen-pairs by $\{\sigma_i, u_i\}$, where the eigenvalues $\{\sigma_i\}$ are non-increasingly ordered and the eigenvectors are subject to normalization $u_i^\top B u_j = \delta_{i,j}$. Thus, we have $\overline{A}_1 = \sum_{i=1}^n \sigma_i u_i u_i^\top$ (assuming that all σ_i's are positive; otherwise, we drop those zero eigenvalues). The least-squares estimators from b and b' are $c = \sum_{i=1}^n \frac{u_i^\top b}{\sigma_i} u_i$ and $c' = \sum_{i=1}^n \frac{u_i^\top b'}{\sigma_i} u_i$, respectively. Third, the difference between their function estimators represents the sampling error (with $\Delta c = c - c'$)

$$\begin{aligned} g(n) :=& \|\widehat{f} - \widehat{f}'\|^2_{L^2(\overline{\rho}_T^L)} = \| \sum_{k=1}^n \Delta c_k \phi_k \|^2_{L^2(\overline{\rho}_T^L)} = \sum_{i,j=1}^n \Delta c_i \langle \phi_i, \phi_j \rangle_{L^2(\overline{\rho}_T^L)} \Delta c_j = \Delta c^\top B \Delta c \\ =& \sum_{i,j=1}^n \frac{u_i^\top (b - b')}{\sigma_i} u_i^\top B u_j \frac{u_j^\top (b - b')}{\sigma_j} = \sum_{i=1}^n r_i^2, \end{aligned} \tag{C.3}$$

where $r_i = \frac{|u_i^\top (b - b')|}{\sigma_i}$. The ratio r_i is in the same spirit as the *Picard projection ratio* $\frac{|u_i^\top b|}{\sigma_i}$ in [36],

which is used to detect overfitting. Note that the eigenvalues σ_i will vanish as n increases because the operator L_{K_1} is compact. Clearly, the sampling error $g(n)$ should be less than $\|f_*\|^2_{L^2(\overline{\rho}_T^L)}$, which is the average of the second moments. Thus, we set N to be

$$N = \max\{k \geq 1 : g(k) \leq \tau\}, \text{ where } \tau = \frac{1}{LM} \sum_{l=1,m=1}^{L,M} |Y_{t_l}^{(m)}|^2. \tag{C.4}$$

We note that this threshold is relatively large, neglecting the rich information in g, a subject worthy of further investigation.

Algorithm 4 summarizes the above procedure.

Input: The state space model and data $\{Y_{t_0:t_L}^{(m)}\}_{m=1}^M$.

Output: A range $[1, N]$ for the dimension of the hypothesis space for further selection.

1: Estimate the empirical density $\overline{\rho}_T^L$ in (4.16) and find its support $[R_{min}, R_{max}]$.

2: Set $n = 1$ and $g(n) = 0$. Estimate the threshold τ in (C.4).

3: **while** $g(n) \leq \tau$ **do**

4: Set $n \leftarrow n + 1$. Update the basis functions, Fourier or B-spline, as in Chapter 4 Section 2.3.

5: Compute normal matrix $\overline{A}_1$ in (4.8) by Monte Carlo. Also, compute b and b' in (C.1).

6: Eigen-decomposition of $\overline{A}_1$ as in (C.2); return $\overline{A}_1 = \sum_{i=1}^n u_i \sigma_i u_i^T$ with $u_i^\top B u_j = \delta_{i,j}$.

7: Compute the Picard projection ratios: $r_i = \frac{|u_i^\top (b - b')|}{\sigma_i}$ for $i = 1, \ldots, n$ and $g(n) = \sum_{i=1}^n r_i^2$.

8: Return $N = n$.

Algorithm 4: Cross-validating Estimation of Dimension Range (CEDR) for hypothesis space

Step 2: We select the dimension n and degree for B-spline basis functions to be the one with the smallest 2-Wasserstein distance between the distribution of the data and that of the predictions. More precisely, let $\mu_{t_l}^f$ and $\mu_{t_l}^{\widehat{f}}$ denote the distributions of $Y_{t_l} = f(X_{t_l})$ and of $\widehat{f}(X_{t_l})$,

respectively. Let F_{t_l} and $\widehat{F}_{t_l}$ denote their cumulative distribution functions (CDF), with $F_{t_l}^{-1}$ and $\widehat{F}_{t_l}^{-1}$ being their corresponding inverses. We compute F_{t_l} from the data and $\widehat{F}_{t_l}$ from independent simulations, approximate their inverses using quantiles, and consider the root mean squared 2-Wasserstein distance

$$\left(\frac{1}{L}\sum_{l=1}^{L} W_2(\mu_{t_l}^f, \mu_{t_l}^{\widehat{f}})^2\right)^{1/2}, \text{ with } W_2(\mu_{t_l}^f, \mu_{t_l}^{\widehat{f}}) = \left(\int_0^1 (F_{t_l}^{-1}(r) - \widehat{F}_{t_l}^{-1}(r))^2 dr\right)^{\frac{1}{2}}. \tag{C.5}$$

This method of computing the Wasserstein distance is based on an observation in [14], and it has been used in [50, 74]. Recall that the 2-Wasserstein distance $W_2(\mu, \nu)$ of two probability measures μ and ν over Ω with finite second order moments is defined as

$$W_2(\mu, \nu) := \inf_{\gamma \in \Gamma(\mu,\nu)} \left(\int_{\Omega\times\Omega} |x - y|^2 d\gamma(x, y)\right)^{1/2},$$

where $\Gamma(\mu, \nu)$ denotes the set of all measures on $\Omega \times \Omega$ with μ and ν as marginals. Let F and G be the CDFs of μ and ν respectively, and let F^{-1} and G^{-1} be their quantile functions. Then the L^2 distance of the quantile functions $d_2(\mu, \nu) := \left(\int_0^1 |F^{-1}(r) - G^{-1}(r)dr|^2\right)^{1/2}$ is equal to the 2-Wasserstein distance $W_2(\mu, \nu)$.

1.2. Optimization with multiple initial conditions With the convex hypothesis space in (4.15), the minimization in (4.12) is a constrained optimization problem and it may have multiple local minima. Note that the loss functional $\mathcal{E}^M$ in (4.12) consists of a quadratic term and two quartic terms. The quadratic term $\mathcal{E}_1^M$ in (4.7), has a Hessian matrix $\overline{A}_1$ which is often not full rank because it is the average of rank-one matrices by its definition in (4.8), in which case the quadratic term has a valley of minima in the kernel of $\overline{A}_1$. The two quartic terms have valleys of

minima at the intersections of the ellipse-shaped manifolds $\{c \in \mathbb{R}^n : c^\top A_{k,l} c = b_{k,l}^M\}_{l=1}^L$ for $k = 2, 3$. Symmetry in the distribution of the state process will also lead to multiple minima (see Section 3.2 for more discussions, and the numerical examples).

To reduce the possibility of obtaining a local minimum, we search for a minimizer from multiple initial conditions. We consider the following initial conditions: (1) the least squares estimator for the quadratic term; (2) the minimizer of the quadratic term in the hypothesis space, which is solved by least squares with linear constraints using ©MATLAB function lsqlin, starting from the LSE estimator; (3) the minimizers of the quartic terms over the hypothesis space, which is found by constrained optimization through ©MATLAB fmincon with the interior-point search. Among the minimizers obtained from these initial conditions, we take the one leading to the smallest 2-Wasserstein distance.

2. Selection of dimension and degree of the B-spline basis

We demonstrate the selection of the dimension and degree of the B-spline basis functions of the hypothesis space. As described in Chapter 4 Section 2.3, we select the dimension and degree in two steps: we first select a rough range for the dimension by the Cross-validating Estimation of Dimension Range (CEDR) algorithm; then we pick the dimension and degree to be the ones with minimal 2-Wasserstein distance between the true and estimated distribution of the observation processes.

The CEDR algorithm helps to reduce the computational cost by estimating the dimension range for the hypothesis space. It is based on an SVD analysis of the normal matrix $\overline{A}_1$ and vector $\overline{b}_1$ from the quadratic loss functional $\mathcal{E}_1$. The key idea is to control the sampling error's effect on the estimator in the metric of the function space of learning. The sampling error is estimated

by computing two copies of the normal vector through splitting the data into two halves. The function space of learning plays an important role here: it directs us to use a generalized eigenvalue problem for the SVD analysis. This is different from the classical SVD analysis in [36], where the information of the function space is neglected.

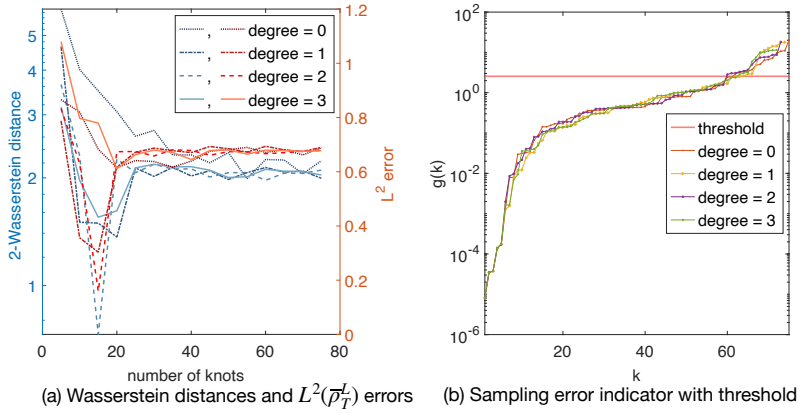


Figure C.1: Selection of the dimension and degree of B-spline basis in the case of Sine-Cosine function. In (a), the 2-Wasserstein distance reaches minimum among all cases when the degree is 2 and the knot number is 15, at the same time as the $L^2(\overline{\rho}_T^L)$ error reaches the minimum. Figure (b) shows the cross-validating error indicator g (defined in (C.3)) for selecting the dimension range N, suggesting an upper bound $N = 60$ with the threshold.

Figure C.2 shows the dimension selection by 2-Wasserstein distances and by the CEDR algorithm for the example of Sine-Cosine function. To confirm the effectiveness of our CEDR algorithm, we compute the 2-Wasserstein distances for all dimensions in (a), side-by-side with the CEDR sampling error indicator g in (b) with relatively large dimensions $\{n = 75 - deg|$ for $deg \in \{0, 1, 2, 3\}$. First, the left figure suggests that the optimal dimension and degree are $n = 13$ and $deg = 2$, where the 2-Wasserstein distance reaches minimum among all cases, and at the same time as the $L^2(\overline{\rho}_T^L)$ error. For the other degrees, the minimum 2-Wasserstein distances are either reached before of after the $L^2(\overline{\rho}_T^L)$ error. Thus, the 2-Wasserstein distance correctly selects the optimal dimension and

degree for the hypothesis space. Second, (a) shows that the CEDR algorithm can effectively select the dimension range. With the threshold in (C.4) being $\tau = 1.60$, which is relatively large (representing a tolerance of 100% relative error), the dimension upper bounds are around $N = 60$ for all degrees, and the ranges encloses the optimal dimensions selected by the 2-Wasserstein distance in (b).

Here we used a relatively large threshold for a rough estimation of the range of dimension. Clearly, our cross-validating error indicator $g(k)$ in (C.3) provides rich information about the increase of sampling error as the dimension increases. A future direction is to extract the information, along with the decay of the integral operator, to control, both in theory and algorithmically, the trade-off between sampling error and approximation error.

Bibliography

[1] A. Agliari and C. C. Parisetti. "A-g Reference Informative Prior: A Note on Zellner's g Prior". In: *Journal of the Royal Statistical Society. Series D (The Statistician)* 37.3 (1988), pp. 271–275.

[2] F. Andreu-Vaillo et al. *Nonlocal Diffusion Problems*. Vol. 165. Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society, 2010. ISBN: 978-0-8218-5230-9 978-1-4704-1392-7.

[3] D. Applebaum. *Lévy processes and stochastic calculus*. Cambridge university press, 2009.

[4] N. Aronszajn. "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3 (1950), pp. 337–404.

[5] P. Bates and A. Chmaj. "An integrodifferential model for phase transitions: stationary solutions in higher space dimensions". In: *Journal of Statistical Physics* 95 (1999), pp. 1119–1139.

[6] M. J. Bayarri et al. "Criteria for Bayesian model choice with application to variable selection". In: *The Annals of Statistics* 40.3 (2012). ISSN: 0090-5364.

[7] D. Benson, S. Wheatcraft, and M. Meerschaert. "Application of a fractional advection-dispersion equation". In: *Water Resources Research* 36.6 (2000), pp. 1403–1412.

[8] C. Berg, J. P. R. Christensen, and P. Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*. Vol. 100. New York: Springer, 1984.

[9] S. A. Billings. *Nonlinear System Identification*. Chichester, UK: John Wiley & Sons, Ltd, 2013. ISBN: 978-1-118-53556-1 978-1-119-94359-4.

[10] P. Brockwell and R. Davis. *Time series: theory and methods*. 2nd. Springer, New York, 1991.

[11] A. Buades, B. Coll, and J. M. Morel. "Image denoising methods: a new nonlocal principle". In: *SIAM Review* 52 (2010), pp. 113–147.

[12] C. Bucur and E. Valdinoci. *Nonlocal Diffusion and Applications*. Vol. 20. Lecture Notes of the Unione Matematica Italiana. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-28738-6 978-3-319-28739-3.

[13] O. Cappé, E. Moulines, and T. Rydén. *Inference in Hidden Markov Models*. Springer Series in Statistics. New York ; London: Springer, 2005. ISBN: 978-0-387-40264-2.

[14] J. A. Carrillo and G. Toscani. "Wasserstein metric and large–time asymptotics of nonlinear diffusion equations". In: *New Trends in Mathematical Physics: In Honour of the Salvatore Rionero 70th Birthday*. World Scientific, 2004, pp. 234–244.

[15] C. K. Chen and P. C. Fife. "Nonlocal Models Of Phase Transitions In Solids". In: *Advances in Mathematical Sciences and Applications* 10.2 (2000), pp. 821–849.

[16] Y. Chen, H. Owhadi, and A. M. Stuart. "Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation". In: *arXiv preprint arXiv:2005.11375* (2021).

[17] Z.-Q. Chen and X. Zhang. "Heat Kernels for Non-Symmetric Non-Local Operators". In: *Recent Developments in Nonlocal Theory*. Ed. by G. Palatucci and T. Kuusi. De Gruyter Open, 2017, pp. 24–51. ISBN: 978-3-11-057156-1.

[18] R. R. Coifman et al. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005), pp. 7426–7431.

[19] F. Cucker and S. Smale. "On the mathematical foundations of learning". In: *Bulletin of the American mathematical society* 39.1 (2002), pp. 1–49.

[20] F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*. Vol. 24. Cambridge University Press, 2007.

[21] K. Dayal and K. Bhattacharya. "Kinetics of phase transformations in the peridynamic formulation of continuum mechanics". In: *Journal of the Mechanics and Physics of Solids* 54.9 (2006), pp. 1811–1842.

[22] M. D'Elia et al. "Numerical Methods for Nonlocal and Fractional Models". In: *Acta Numerica* 29 (2020), pp. 1–124. ISSN: 0962-4929, 1474-0508.

[23] Q. Du et al. "Analysis and Approximation of Nonlocal Diffusion Problems with Volume Constraints". In: *SIAM Rev.* 54.4 (2012), pp. 667–696. ISSN: 0036-1445, 1095-7200.

[24] L. C. Evans. *Partial differential equations*. Vol. 19. American Mathematical Soc., 2010.

[25] J. Fan and Q. Yao. *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York, NY, 2003.

[26] R. D. Fierro et al. "Regularization by Truncated Total Least Squares". In: *SIAM J. Sci. Comput.* 18.4 (1997), pp. 1223–1241. ISSN: 1064-8275, 1095-7197.

[27] C. Gelada et al. "DeepMDP: Learning Continuous Latent Space Models for Representation Learning". In: *arXiv preprint arXiv:1906.02736* (2019).

[28] A. Ghosh et al. "Bayesian Inference in Nonparametric Dynamic State-Space Models". In: *Statistical Methodology* 21 (2014), pp. 35–48. ISSN: 15723127.

[29] G. Gilboa and S. Osher. "Nonlocal linear image regularization and supervised segmentation". In: *Multiscale Modeling and Simulation* 6 (2007), pp. 595–630.

[30] G. Gilboa and S. Osher. "Nonlocal operators with applications to image processing". In: *Multiscale Modeling & Simulation* 7.3 (2009), pp. 1005–1028.

[31] C. R. Gin et al. "DeepGreen: deep learning of Green's functions for nonlinear boundary value problems". In: *Scientific reports* 11.1 (2021), pp. 1–14.

[32] N. Guglielmi and E. Hairer. "Classification of Hidden Dynamics in Discontinuous Dynamical Systems". In: *SIAM J. Appl. Dyn. Syst.* 14.3 (2015), pp. 1454–1477. ISSN: 1536-0040.

[33] L. Györfi et al. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[34] D. Hafner et al. "Learning Latent Dynamics for Planning from Pixels". In: *arXiv preprint arXiv:1811.04551* (2019).

[35] B. Hamzi and H. Owhadi. "Learning Dynamical Systems from Data: A Simple Cross-Validation Perspective, Part I: Parametric Kernel Flows". In: *Physica D: Nonlinear Phenomena* 421 (2021), p. 132817. ISSN: 01672789.

[36] P. C. Hansen. "The L-Curve and its Use in the Numerical Treatment of Inverse Problems". In: (2000), pp. 119–142.

[37] P. C. Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

[38] G. Holler and K. Kunisch. "Learning Nonlocal Regularization Operators". In: *arXiv preprint arXiv:2001.09092* (2020).

[39] T. Hsing and R. Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*. Vol. 997. John Wiley & Sons, 2015.

[40] R. Hwang et al. "Solving PDE-constrained Control Problems using Operator Learning". In: *arXiv preprint arXiv:2111.04941* (2021).

[41] P.-E. Jabin and Z. Wang. "Mean Field Limit for Stochastic Particle Systems". In: *Active Particles, Volume 1*. Springer, 2017, pp. 379–402.

[42] M. R. Jeffrey. *Hidden Dynamics: The Mathematics of Switches, Decisions and Other Discontinuous Behaviour*. Cham: Springer International Publishing, 2018. ISBN: 978-3-030-02106-1 978-3-030-02107-8.

[43] B. Jin and W. Rundell. "A tutorial on inverse problems for anomalous diffusion processes". In: *Inverse problems* 31.3 (2015), p. 035003.

[44] H. Kadri et al. "Operator-Valued Kernels for Learning from Functional Response Data". In: *J. Mach. Learn. Res.* 17.1 (2016), 613–666. ISSN: 1532-4435.

[45] Łukasz Kaiser et al. "Model Based Reinforcement Learning for Atari". In: *International Conference on Learning Representations*. 2020.

[46] N. Kantas et al. "An Overview of Sequential Monte Carlo Methods for Parameter Estimation in General State-Space Models". In: *IFAC Proc. Vol.* 42.10 (2009), pp. 774–785. ISSN: 14746670.

[47] A. Katiyar et al. "A general peridynamics model for multiphase transport of non-Newtonian compressible fluids in porous media". In: *Journal of Computational Physics* (2019). In press.

[48] A. Katiyar et al. "A peridynamic formulation of pressure driven convective fluid transport in porous media". In: *Journal of Computational Physics* 261 (2014), pp. 209–229.

[49] S. Kindermann, S. Osher, and P. W. Jones. "Deblurring and denoising of images by nonlocal functionals". In: *Multiscale Modeling & Simulation* 4.4 (2005), pp. 1091–1115.

[50] N. Kolbe. *Wasserstein Distance*. `https://github.com/nklb/wasserstein-distance`. 2020.

[51] N. Kovachki et al. "Neural operator: Learning maps between function spaces". In: *arXiv preprint arXiv:2108.08481* (2021).

[52] Q. Lang and F. Lu. "Identifiability of interaction kernels in mean-field equations of interacting particles". In: *arXiv preprint arXiv:2106.05565* (2021).

[53] Q. Lang and F. Lu. "Learning interaction kernels in mean-field equations of 1st-order systems of interacting particles". In: *arXiv2010.15694* (2020).

[54] K. Law, A. Stuart, and K. Zygalakis. *Data Assimilation: A Mathematical Introduction*. Springer, 2015.

[55] Y. Li et al. "Extracting stochastic dynamical systems with *alpha*-stable Lévy noise from data". In: *arXiv preprint arXiv:2109.14881* (2021).

[56] Z. Li et al. "On the identifiability of interaction functions in systems of interacting particles". In: *Stochastic Processes and their Applications* 132 (2021), pp. 135–163.

[57] Z. Li et al. "Fourier Neural Operator for Parametric Partial Differential Equations". In: *International Conference on Learning Representations*. 2020.

[58] Z. Li et al. "Neural operator: Graph kernel network for partial differential equations". In: *arXiv preprint arXiv:2003.03485* (2020).

[59] Z. Li et al. "Neural Operator: Graph Kernel Network for Partial Differential Equations". In: *arXiv preprint arXiv:2003.03485* (2020).

[60] C. Lin et al. "A seamless multiscale operator neural network for inferring bubble dynamics". In: *Journal of Fluid Mechanics* 929 (2021).

[61] C. Lin et al. "Operator learning for predicting multiscale bubble growth dynamics". In: *The Journal of Chemical Physics* 154.10 (2021), p. 104118.

[62] L. Ljung. "System identification". In: *Signal analysis and prediction*. Springer, 1998, pp. 163–173.

[63] H. J. Logarzo, G. Capuano, and J. J. Rimoli. "Smart constitutive laws: Inelastic homogenization through machine learning". In: *Computer Methods in Applied Mechanics and Engineering* 373 (2021), p. 113482. ISSN: 0045-7825.

[64] Y. Lou et al. "Image recovery via nonlocal operators". In: *Journal of Scientific Computing* 42 (2010), pp. 185–197.

[65] F. Lu, Q. Lang, and Q. An. "Data adaptive RKHS Tikhonov regularization for learning kernels in operators". In: *arXiv preprint arXiv:2203.03791* (2022).

[66] F. Lu, M. Maggioni, and S. Tang. "Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories". In: *Foundations of Computational Mathematics* (2021), pp. 1–55.

[67] F. Lu et al. "Nonparametric inference of interaction laws in systems of agents from trajectory data". In: *Proc. Natl. Acad. Sci. USA* 116.29 (2019), pp. 14424–14433.

[68] L. Lu, P. Jin, and G. E. Karniadakis. "Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators". In: *arXiv preprint arXiv:1910.03193* (2019).

[69] L. Lu et al. "Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators". In: *Nature Machine Intelligence* 3.3 (2021), pp. 218–229.

[70] T. Lyche, C. Manni, and H. Speleers. "Foundations of Spline Theory: B-Splines, Spline Approximation, and Hierarchical Refinement". In: *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra*. Vol. 2219. Cham: Springer International Publishing, 2018, pp. 1–76. ISBN: 978-3-319-94910-9 978-3-319-94911-6.

[71] C. Moosmüller, F. Dietrich, and I. G. Kevrekidis. "A Geometric Approach to the Transport of Discontinuous Densities". In: *arXiv preprint arXiv:1907.08260* (2019).

[72] S. Mostch and E. Tadmor. "Heterophilious Dynamics Enhances Consensus". In: *SIAM Rev* 56.4 (2014), pp. 577 –621.

[73] H. Owhadi and G. R. Yoo. "Kernel Flows: From Learning Kernels from Data into the Abyss". In: *Journal of Computational Physics* 389 (2019), pp. 22–47. ISSN: 00219991.

[74] V. M. Panaretos and Y. Zemel. "Statistical aspects of Wasserstein distances". In: *Annual review of statistics and its application* 6 (2019), pp. 405–431.

[75] L. Piegl and W. Tiller. *The NURBS Book*. Monographs in Visual Communication. Berlin, Heidelberg: Springer Berlin Heidelberg, 1997. ISBN: 978-3-540-61545-3 978-3-642-59223-2.

[76] Y. Pokern, A. M. Stuart, and P. Wiberg. "Parameter Estimation for Partially Observed Hypoelliptic Diffusions". In: *J. R. Stat. Soc. Ser. B Stat. Methodol.* 71.1 (2009), pp. 49–73. ISSN: 13697412, 14679868.

[77] B. L. S. Prakasa Rao. "Statistical Inference from Sampled Data for Stochastic Processes". In: *Contemporary Mathematics*. Ed. by N. U. Prabhu. Vol. 80. Providence, Rhode Island: American Mathematical Society, 1988, pp. 249–284. ISBN: 978-0-8218-5087-9 978-0-8218-7668-8.

[78] A. Rahimi and B. Recht. "Unsupervised regression with applications to nonlinear system identification". In: *Advances in Neural Information Processing Systems*. 2007, pp. 1113–1120.

[79] G. Raskutti, M. J. Wainwright, and B. Yu. "Early stopping and non-parametric regression: an optimal data-dependent stopping rule". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 335–366.

[80] R. Schumer et al. "Eulerian derivation of the fractional advection-dispersion equation". In: *Journal of Contaminant Hydrology* 48 (2001), pp. 69–88.

[81] R. Schumer et al. "Multiscaling fractional advection-dispersion equations and their solutions". In: *Water Resources Research* 39.1 (2003), pp. 1022–1032.

[82] L. Shi, Y.-L. Feng, and D.-X. Zhou. "Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces". In: *Applied and Computational Harmonic Analysis* 31.2 (2011), pp. 286–302.

[83] S. A. Silling. "Propagation of a Stress Pulse in a Heterogeneous Elastic Bar". In: *Sandia Report SAND2020-8197, Sandia National Laboratories* (2020).

[84] S. A. Silling et al. "Peridynamic States and Constitutive Modeling". In: *J Elasticity* 88.2 (2007), pp. 151–184. ISSN: 0374-3535, 1573-2681.

[85] M. Sørensen. "Estimating Functions for Diffusion-Type Processes". In: *Statistical Methods for Stochastic Differential Equations*. Vol. 124. Monogr. Statist. Appl. Probab, 2012, pp. 1–107.

[86] H. Sun. "Mercer theorem for RKHS on noncompact sets". In: *Journal of Complexity* 21.3 (2005), pp. 337 –349. ISSN: 0885-064X.

[87] A. Svensson and T. B. Schön. "A Flexible State–Space Model for Learning Nonlinear Dynamical Systems". In: *Automatica* 80 (2017), pp. 189–199. ISSN: 00051098.

[88] A. N. Tihonov. "Solution of incorrectly formulated problems and the regularization method". In: *Soviet Math.* 4 (1963), pp. 1035–1038.

[89] F. Tobar, P. M. Djuric, and D. P. Mandic. "Unsupervised State-Space Modeling Using Reproducing Kernels". In: *IEEE Trans. Signal Process.* 63.19 (2015), pp. 5210–5221. ISSN: 1053-587X, 1941-0476.

[90] H. Wang and T. S. Basu. "A fast finite difference method for two-dimensional space-fractional diffusion equations". In: *SIAM Journal on Scientific Computing* 34.5 (2012), A2444–A2458.

[91] H. Wang. "Analysis of statistical learning algorithms in data dependent function spaces". PhD thesis. City University of Hong Kong, 2009.

[92] X. Wang et al. "Non-Local Neural Networks". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, 2018, pp. 7794–7803. ISBN: 978-1-5386-6420-9.

[93] J. Xiong, J. Zheng, and X. Zhou. "Unique strong solutions of Lévy processes driven stochastic differential equations with discontinuous coefficients". In: *Stochastics* 91.4 (2019), pp. 592–612.

[94] F. X. F. Ye, S. Yang, and M. Maggioni. *Nonlinear model reduction for slow-fast stochastic systems near manifolds*. 2021.

[95] H. You et al. "A data-driven peridynamic continuum model for upscaling molecular dynamics". In: *Computer Methods in Applied Mechanics and Engineering* 389 (2022), p. 114400.

[96] H. You et al. "Data-Driven Learning of Nonlocal Models: From High-Fidelity Simulations to Constitutive Laws". In: *arXiv preprint arXiv:2012.04157* (2020).

[97] H. You et al. “Data-Driven Learning of Nonlocal Physics from High-Fidelity Synthetic Data”. In: *Computer Methods in Applied Mechanics and Engineering* 374 (2021), p. 113553. ISSN: 00457825.

[98] H. You et al. “Nonlocal Kernel Network (NKN): a Stable and Resolution-Independent Deep Neural Network”. In: *arXiv preprint arXiv:2201.02217* (2022).

[99] A. Zellner and A. Siow. “Posterior odds ratios for selected regression hypotheses”. In: *Trabajos de Estadistica Y de Investigacion Operativa* 31 (1980), pp. 585–603.