Proceedings of Machine Learning Research vol 145:1–23, 2022 3rd Annual Conference on Mathematical and Scientific Machine Learning

# Data adaptive RKHS Tikhonov regularization for learning kernels in operators

**Fei Lu** FEILU@MATH.JHU.EDU
**Quanjun Lang** QLANG1@JHU.EDU
**Qingci An** QAN2@JHU.EDU
*Department of Mathematics, Johns Hopkins University, Baltimore, MD*

**Editors:** Bin Dong, Qianxiao Li, Lei Wang, Zhi-Qin John Xu

## Abstract

We present DARTR: a Data Adaptive RKHS Tikhonov Regularization method for the linear inverse problem of nonparametric learning of function parameters in operators. A key ingredient is a system intrinsic data adaptive (SIDA) RKHS, whose norm restricts the learning to take place in the function space of identifiability. DARTR utilizes this norm and selects the regularization parameter by the L-curve method. We illustrate its performance in examples including integral operators, nonlinear operators and nonlocal operators with discrete synthetic data. Numerical results show that DARTR leads to an accurate estimator robust to both numerical error and measurement noise, and the estimator converges at a consistent rate as the data mesh refines under different levels of noises, outperforming two baseline regularizers using $l^2$ and $L^2$ norms.

**Keywords:** ill-posed inverse problem, Tikhonov regularization, RKHS, identifiability

## 1. Introduction

Regularization plays a crucial role in machine learning and inverse problems that aim to construct robust generalizable models. The learning of kernel functions in operators is such a problem: given data consisting of discrete noisy observations of function pairs $\{(u_k, f_k)\}_{k=1}^N$, we would like to learn an optimal kernel function $\phi$ fitting the operator $R_\phi[u] = f$ to the data. Such a need for learning operators between function spaces has become vital in applications ranging from integral operators solving PDEs and image processing (see e.g., Gin et al. (2021); Li et al. (2020); Kovachki et al. (2021); Owhadi and Yoo (2019)), nonlinear operators in mean-field equation of interacting particle systems in Lu et al. (2021); Lang and Lu (2022), homogenized nonlocal operators (see e.g., You et al. (2021, 2022); Lin et al. (2021)), 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 the data mesh refines and is robust to noise in data. Without proper regularization, the estimator often oscillates largely with datasets due to overfitting. Thus, regularization is crucial for the discovery of a generalizable kernel.

We present DARTR, 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 $\phi$. 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. In numerical examples including integral operators, nonlinear operators and nonlocal operators with discrete noisy synthetic data, DARTR consistently leads to accurate estimators, and the estimator

converges at a consistent rate as the data mesh refines under different levels of noises, outperforming two baseline regularizers using $l^2$ and $L^2$ norms.

The major novelty of this method is the construction of a system (the operator) intrinsic data adaptive (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 learning to take place in the function space of identifiability. 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.

### 1.1. Related work

**Classical regression.** The function space of identifiability (FSOI) is a fundamental difference between the classical regression and the regression of kernels in operators. Here the FSOI must be specified properly, otherwise the inverse problem can be *ill-defined* in the sense that there are multiple kernels fitting the data. In contrast, the classical regression inversion is always well-defined and the conditional mean is the unique minimizer. More specifically, the classical regression learns a function $Y = \phi(X)$ from random samples $\{(X_i, Y_i)\}$, its FSOI is $L^2(\rho)$ with $\rho$ being the distribution of $X$, and the optimal estimator is $\mathbb{E}[Y|X]$, see e.g., Cucker and Smale (2002); Györfi et al. (2006)). It corresponds to our setting with $R_\phi(u) = \phi(u)$ with data $\{(u_i, f_i) = (X_i, Y_i)\}$ (i.e., $R_\phi$ is not an operator but a function), and our DARTR reduces to the $L^2$ Tikhonov/ridge regularization (see e.g., Tihonov (1963)).

**Functional data analysis (FDA):** Learning kernels in operators falls in the category of FDA that learns an operator or a functional. Different from those assuming no structure of the operator Hsing and Eubank (2015); Kadri et al. (2016); Ferraty and Vieu (2006), we exploit a low-dimensional structure of the operator: a radial kernel, which can be viewed as a function parameter, thus, we can learn the kernel (and hence the operator) from only a few pairs of data. The setting of linear operators in our study is similar to the functional linear models (FLM) (see e.g., Ramsay and Silverman (2005) and Wang et al. (2016)), for which regression and regularization are also used. In comparison with the regularizations for these FLMs that are based on extra differential operators based prior assumptions, the major novelty of our DARTR is its utilization of a SIDA-RKHS based on our new identifiability theory. Also, our method and theory are applicable to nonlinear operators.

**Tikhonov regularization methods.** DARTR differs from other Tikhonov/ridge regularization methods at the penalty term. The commonly used penalty terms include the Euclidean norm in the classical Tikhonov regularization (see e.g., Hansen (1994, 2000); Gazzola et al. (2019)), the RKHS norm with an ad hoc reproducing kernel (see e.g., Cucker and Zhou (2007); Bauer et al. (2007)), the total variation norm in the Rudin-Osher-Fatemi method in Rudin et al. (1992), or the $L^1$ norm in LASSO (see e.g., Tibshirani (1996)). Whereas each of them has specific applications, none of them take into account of the FSOI, which is fundamental for the learning of kernels in operators.

**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 in Wang (2009); Shi et al. (2011) and data-dependent early stopping rule in Raskutti et al. (2014). While all strategies achieve data-dependent regularization, only our DARTR takes into account the function space of identifiably.

## 2. The inverse problem and the need of regularization

### 2.1. Problem statement: learning function parameters in operators

We consider the linear inverse problem of identifying function parameters in operators from data, that is, to learn a function parameter $\phi$ in ab operator $R_\phi : \mathbb{X} \to \mathbb{Y}$ of the form $R_\phi[u] = f$ from data pairs $\{(u_k, f_k)\}_{k=1}^N \subset \mathbb{X} \times \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are problem specific functions spaces. Specifically, suppose that we are given data consisting of discrete observation of function pairs:

$$\mathcal{D} = \{(u_k, f_k)\}_{k=1}^N = \{(u_k(x_j), f_k(x_j)) : j = 1, \dots, J\}_{k=1}^N, \tag{2.1}$$

where $(u_k, f_k)$ are real-valued functions on a bounded open set $\Omega \subset \mathbb{R}^d$ and $\{x_j \in \Omega\}$ are spatial mesh points. For simplicity, we assume $\mathbb{Y} = L^2(\Omega)$ and assume $\mathbb{X}$ to be operator specific. Our goal is to learn $\phi$ in an operator $R_\phi$ fitting the data in the form:

$$R_\phi[u](x) = \int_\Omega \phi(|y|) g[u](x, y) dy, \ \forall x \in \Omega, \tag{2.2}$$

where the functional $g$, which may depend on the derivatives of $u$, is known and it specifies the form of the operator. Examples are as follows (see more details in Section 5):

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

These inverse problems share three common features: First, the pointwise values of the function $\phi$ are undetermined from data, because the data depends on $\phi$ non-locally. Also, the support of $\phi$ is unknown and is to be learnt from data. Second, the data are discrete and can have measurement noise. Thus, the inverse problem has to overcome the numerical error in the approximation of integrals, as well as the noise. Third, the inverse problem can be extended to a homogenization problem where the operator aims to fit the data that are not generated from the equation (2.2). In this case, the inverse problem has to overcome the model error to identify a best fit.

### 2.2. Nonparametric regression and regularization

Our goal is to infer the kernel function $\phi$ from data in a nonparametric fashion, so as to address the general situations that there is limited information to derive a parametric form for the kernel. Thus, we will not assume any constraint on the function $\phi$. More importantly, we aim for an estimator that is consistent and resolution independent, i.e., converges in a proper function space to the true kernel as data mesh refines and is robust to treat noisy data.

We construct a variational estimator that minimizes loss functional (the mean square error),

$$\widehat{\phi} = \arg\min_{\phi \in \mathcal{H}} \mathcal{E}(\phi), \quad \text{where } \mathcal{E}(\phi) = \frac{1}{N} \sum_{k=1}^N \|R_\phi[u_k] - f_k\|_{\mathbb{Y}}^2, \tag{2.3}$$

where the hypothesis space $\mathcal{H}$ is to be selected adaptive to data. Note that the loss functional $\mathcal{E}(\phi)$ is quadratic in $\phi$ since the operator $R_\phi$ depends linearly on $\phi$. Thus, the minimizer of the loss functional is a least squares estimator. Suppose the hypothesis space is $\mathcal{H}_n = \text{span}\{\phi_i\}_{i=1}^n$ with basis functions $\{\phi_i\}$. 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 $\mathcal{E}(c) = \mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^f$, where $C_N^f = \frac{1}{N}\sum_{k=1}^N \int_\Omega |f_k(x)|^2 dx$ and the normal matrix $\overline{A}_n$ and vector $\overline{b}_n$ are given by

$$\overline{A}_n(i,j) = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle, \quad \overline{b}_n(i) = \frac{1}{N}\sum_{k=1}^N \langle R_{\phi_i}[u_k], f_k\rangle_{\mathbb{Y}}, \tag{2.4}$$

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

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \frac{1}{N}\sum_{k=1}^N \langle R_\phi[u_k], R_\psi[u_k]\rangle_{\mathbb{Y}}. \tag{2.5}$$

The least squares estimator is minimizes the quadratic loss function $\mathcal{E}(c)$:

$$\widehat{\phi}_{\mathcal{H}_n} = \sum_{i=1}^n \widehat{c}_i \phi_i \quad \text{and} \quad \widehat{c} = \overline{A}_n^{-1}\overline{b}_n, \tag{2.6}$$

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 to imperfect data and model error, in particular, converging in synthetic tests when the data mesh refines. Unfortunately, this is an ill-posed inverse problem (see Section 3.2) and the normal matrix $\overline{A}_n$ is often highly ill-conditioned or singular. As a result, the estimator in (2.6) oscillates largely and fails to converge when the data mesh refines.

Various regularization methods have been introduced to prevent over-fitting in such ill-posed inverse problems. The idea is to add a penalty term to the loss functional:

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \mathcal{R}(\phi) \tag{2.7}$$

where $\mathcal{R}(\phi)$ is a regularization term and $\lambda$ is a parameter that controls the importance of the regularization. Various penalty terms have been proposed, including, for example, the Euclidean norm $\mathcal{R}(\phi) = \|c\|^2$ for $\phi = \sum_{i=1}^n c_i\phi_i$ in the classical Tikhonov regularization (see e.g., Tihonov (1963); Hansen (1998)), the RKHS norm $\mathcal{R}(\phi) = \|\phi\|_H^2$ with $H$ being a reproducing kernel Hilbert space with an artificial reproducing kernel (see e.g., Cucker and Zhou (2007); Bauer et al. (2007)) , 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 (see e.g., Tibshirani (1996)).

Whereas each of these penalty terms has a specific premise for a class of applications, none of them take into account of the function space of identifiability (see Section 3.2), only in which the inverse problem is well-defined. Our DARTR method (see Section 4.1) will utilize a norm that restricts the learning in the function space of identifiability, thus providing a crucial regularization.

## 3. Identifiability theory and regularization

The foundation of learning is the function space of identifiability (FSOI), in which the inverse problem well-defined. We provide a full characterization of it and importantly, we show that it can be a proper subspace of the $L^2$ space, the default function space of learning. Thus, it is vital to restrict the learning to take place in the FSOI. We show that the norm of a system intrinsic data adaptive (SIDA) RKHS, when used for regularization, can achieve the goal.

The main theme of the identifiability theory is 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 (2.5), we can rewrite the loss functional in (2.3) as

$$\mathcal{E}(\phi) = \langle\!\langle \phi, \phi \rangle\!\rangle - 2\frac{1}{N}\sum_{k=1}^{N} \langle R_\phi[u_k], f_k \rangle_{\mathbb{Y}} + C_f, \tag{3.1}$$

where $C_N^f = \frac{1}{N}\sum_{k=1}^{N}\int |f_k(x)|^2 dx$. However, there is no function space for $\phi$ yet.

To start, we introduce two key elements: an exploration measure that leads to a default function space of learning and an integral operator which plays a crucial role in the identifiability theory. Here we consider only the functions $\{u_k, f_k\}$ to simplify the notation. The integrals will be numerically approximated from the discrete data in computation. Also, all the results extend directly to a version on discrete vector space for discrete data, see Remark 3.4.

**Assumption 3.1** *The functions $\{u_k\}_{k=1}^N$ in (2.1) satisfy that each function $g[u_k] : \overline{\Omega} \times \overline{\Omega} \to \mathbb{R}$, which defines the operator in (2.2), is continuous.*

### 3.1. An integral operator and the SIDA-RKHS

**The exploration measure.** We introduce first a probability measure that quantifies the exploration of the variable of $\phi$ by the data. Given data in (2.1), 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. This measure reflects the weight being put on $|y|$ by the loss function through the data $\{g[u_k](x,y)\}_{k=1}^N$.

The exploration measure plays an important role in the learning of the function $\phi$. 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 $\phi$. 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 (2.5). 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. The last equality in (3.3) is a re-weighting by $\rho$ with

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

where, abusing the notation, we use $\rho(r)$ to denote the density of the measure $\rho$ defined in (3.2).

The next lemma shows that $\overline{G}$ defines a positive semi-definite integral operator. Its proof, as well as proofs to later lemmas and theorems, are presented in Appendix A.

**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 $\phi, \psi \in L^2(\rho)$,*

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

The next lemma provides an operator characterization of the RKHS with $\overline{G}$ as the reproducing kernel Aronszajn (1950). This RKHS is system(the operator $R_\phi$) 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 $\lambda_i$ converges to 0. Furthermore, for any $\phi = \sum_i c_i\psi_i$, we have*

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

*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*

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

**Remark 3.4** *The space $L^2(\rho)$ can be a discrete vector space with the function $\phi$ 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. These discrete values can be viewed piecewise constant approximations to the functions, and the numerical algorithm in Section 4.1 applies. When $n \to \infty$, they converge to the corresponding functions under suitable regularity conditions. Thus, the measure $\rho$ allows for a unified framework to treat the SIDA-RKHS with either discrete or continuous functions.*

### 3.2. Function space of identifiability and regularizations

We show that the function space of identifiability (FSOI), i.e., on which the loss functional has a unique minimizer (see Definition 3.5), is the subspace of $L^2(\rho)$ spanned by the eigenfunctions of $\mathcal{L}_{\overline{G}}$ with positive eigenvalues. 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 this function space. Furthermore, we show that the regularization by the SIDA-RKHS norm enforces the regularized minimizer to be in it.

**Definition 3.5** *The function space of identifiability is the largest linear subspace of $L^2(\rho)$ in which the loss functional $\mathcal{E}$ has a unique minimizer.*

We remark that when the data is continuous and noiseless, the true kernel is the unique minimizer, thus, it is identifiable by the loss functional. Also, note that the FSOI is data-dependent. When the data is discrete or noisy, the unique minimizer in the FSOI is an optimal estimator in the FSOI, and it converges to the true kernel as the data improves.

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

**Theorem 3.6 (Function space of identifiability (FSOI))** *Suppose that Assumption 3.1 holds. Let $\phi_N^f \in L^2(\rho)$ be the Riesz representation of the bounded linear functional:*

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

*Then the following statements hold.*

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

*(b) The FSOI 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 $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^f$ if $\phi_N^f \in \mathcal{L}_{\overline{G}}(L^2(\rho))$. In particular, if the true function is $\phi_{true} \in H$ and the data is continuous noiseless, we have $\phi_N^f = \mathcal{L}_{\overline{G}}\phi_{true}$ and $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^f = \phi_{true}$.*

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

**Remark 3.7 (Regularization with the $L^2$ and the SIDA-RKHS norms)** *In practice, due to the discrete and/or noisy data, we often have $\phi_N^f = \mathcal{L}_{\overline{G}}\phi_{true} + \phi_1^\delta + \phi_2^\delta$, 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 $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi_N^f$ does not exist and regularization is necessary. The following comparison between the $L^2$ and the SIDA-RKHS regularizers shows that the later regularizer removes $\phi_2^\delta$ and makes the estimator well-defined. More specifically, we 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*

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

*Plugging in $\phi_N^f = \mathcal{L}_{\overline{G}}\phi_{true} + \phi_1^\delta + \phi_2^\delta$, 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 $\lambda$ to balance the errors,*

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

*where in each of them, the first term on the right hand side requires a large $\lambda$, whereas the second term requires a small $\lambda$. In practice, the errors $\phi_i^\delta$ are much smaller than $\phi_{true}$, and the optimal*

*$\lambda$ should be small so that the second term 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, the SIDA-RKHS regularized estimator $\widehat{\phi}_\lambda^{H_G}$ is more accurate than the $L^2$ regularized estimator. To avoid amplifying the error $\phi_2^\delta$, a projection is necessary for the $L^2$ regularizer, and we will compare the projected $L^2$ regularizer with the SIDA-RKHS regularizer in Section 4.2.*

## 4. Learning algorithm

### 4.1. Algorithm: nonparametric regression with DARTR

Based on the identifiability theory in Section 3.2, 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 $\rho$. We utilize the data to estimate the support of the true kernel and the exploration measure $\rho$. The support of the true kernel lies in $[0, R_0]$ with $R_0$ being the diameter of the domain $\Omega$, and it is further confined from a comparison between the supports of $f_k$ and $g[u_k]$ (see Appendix B.1 for more details). Then, we constrain the discrete approximation of $\rho$ defined (3.2) on the support of $\phi$. In this process, we also assemble the regression data that will be repeatedly used.
2. Assemble the regression matrices and vectors. We select a class of hypothesis spaces $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$ 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,

$$\overline{A}_n(i,j) = \langle\!\langle \phi_i, \phi_j \rangle\!\rangle, \quad \overline{b}_n(i) = \langle \phi_N^f, \phi_i \rangle_{L^2(\rho)}, \quad B_n(i,j) = \langle \phi_i, \phi_j \rangle_{L^2(\rho)}. \tag{4.1}$$

   from data for each of these hypothesis spaces.
3. For each triplet $(\overline{A}_n, \overline{b}_n, B_n)$, find the best regularized estimator $\widehat{c}_{\lambda_n}$ by DARTR in Algorithm 1, as well as corresponding loss value $\mathcal{E}(\widehat{c}_{\lambda_n})$.
4. From the estimators $\{\widehat{c}_{\lambda_n}\}_n$, we select the one with the smallest loss value $\mathcal{E}(\widehat{c}_{\lambda_n})$.

**Algorithm 1** Data Adaptive RKHS Regularization (DARTR).

**Input:** The regression triplet $(\overline{A}, \overline{b}, B)$ consisting of normal matrix $\overline{A}$, vector $\overline{b}$ and basis matrix $B$ as in (4.1).
**Output:** SIDA-RKHS regularized estimator $\widehat{c}_{\lambda_0}$ and loss value $\mathcal{E}(\widehat{c}_{\lambda_0})$.

1: Solve the generalized eigenvalue problem $\overline{A}V = BV\Lambda$, where $\Lambda$ is the diagonal matrix of eigenvalues and the matrix $V$ has columns being eigenvectors orthonormal in the sense that $V^\top BV = I$.
2: Compute the RKHS-norm matrix $B_{rkhs} = (V\Lambda V^\top)^{-1}$, using pseudo inverse when $\Lambda$ is singular. We refer to 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 $\lambda_0$ maximizing the curvature of the $\lambda$-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.

In comparison to the classical nonparametric regression using $(\overline{A}_n, \overline{b}_n)$, we need only an additional basis matrix $B_n$. The novelty of our algorithm is the data adaptive components, such as

the exploration measure $\rho$, 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 $\widehat{\psi}_k = V_{jk}\phi_j$ approximates the eigenfunctions of $\mathcal{L}_{\overline{G}}$ (see Theorem 4.1). The additional computational cost is only the generalized eigenvalue problem which can be solved efficiently.

**Theorem 4.1** *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 (4.1). 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 $\rho$ is a discrete-measure on a discrete set $\mathcal{R}$ and $\{\phi_n\}$ are piecewise constant functions with $n = |\mathcal{R}|$). Then, the operator $\mathcal{L}_{\overline{G}}$ in (3.6) has eigenvalues $(\lambda_1, \ldots, \lambda_n)$ solved by the generalize eigenvalue problem*

$$\overline{A}_n V = B_n \Lambda V, \quad s.t., V^\top B_n V = I_n, \quad \Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_n), \tag{4.2}$$

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

### 4.2. 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 learning to take place in the function space of identifiability. In the following, we compare it with the $l^2$ and $L^2$ regularizers with $\mathcal{R}(\phi) = \|\phi\|_{l^2}^2 = \sum_i c_i^2$ and $\mathcal{R}(\phi) = \|\phi\|_{L^2}^2 = c^\top B_n c$. 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 $L^2(\rho)$ (see also Remark 3.7). Thus, in practice, one makes a projection to the FSOI (i.e., the eigen-space of nonzero eigenvalues of $\overline{A}_n$ in computation) before adding these regularization terms, and we call them projected $l^2$ and $L^2$ regularizers.

Table 1: The SIDA-RKHS regularizer v.s. the projected $l^2$,$L^2$ regularizers*.

| | $l^2$ | $L^2$ | SIDA-RKHS |
|---|---|---|---|
| $\mathcal{R}(\phi)$ | $\|c\|^2 = c^\top c$ | $\|c\|_{B_n}^2 = c^\top B_n c$ | $\|c\|_{H_G}^2 = c^\top B_{rkhs} c$ |
| $c_\lambda$ | $c_\lambda = \sum_{i=1}^k \frac{1}{\sigma_i + \lambda} u_i^\top \overline{b}$ | $c_\lambda = \sum_{i=1}^k \frac{1}{\lambda_i + \lambda} v_i^\top \overline{b}$ | $c_\lambda = \sum_{i=1}^k \frac{1}{\lambda_i + \lambda\lambda_i^{-1}} v_i^\top \overline{b}$ |
| SVD | $\overline{A}_n = \sum_{i=1}^n \sigma_i u_i u_i^\top, u_i^\top u_j = \delta_{ij}$ <br> $U^\top \overline{A}_n U = \Sigma, U^\top U = I$ | $\overline{A}_n = \sum_{i=1}^n \lambda_i v_i v_i^\top, \ v_i^\top B_n v_j = \delta_{ij}$ <br> $V^\top \overline{A}_n V = \Lambda, V^\top B_n V = I$ | |

*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 (4.1)). The projected $l^2$ and $L^2$ regularizers use only the non-zero eigenvalues $\{\sigma_i\}_{i=1}^k$ and $\{\lambda_i\}_{i=1}^k$ and their eigenvectors.

Table 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 or when $\overline{A}_n = B_n$ as in classical

regression), the two are the same. Otherwise, the SIDA-RKHS regularizer not only takes into account of the FSOI but also a balance between $\lambda_i$ and $\lambda_i^{-1}$.

- The SIDA-RKHS regularizer restricts the learning to be in the FSOI by definition, while the other two regularizers, if not projected, miss this fundamental issue.

## 5. Numerical results

We test our learning method on three types of operators: linear integral operators, nonlocal operators and nonlinear operators. For each type, 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.

### 5.1. Settings and main results

The settings of the numerical tests for all three types of operators are as below.

**Comparison with baseline methods.** On each dataset, we compare our SIDA-RKHS regularizer with two baseline regularizers using the projected $l^2$ and $L^2$ norm (denoted by l2 and L2 in the figures below, respectively) defined in Section 4.2. All three regularizers use the same L-curve method to select the hyper-parameter $\lambda$ as described in Appendix B.1. They differ only at the regularization norm.

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

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

• *Gaussian kernel.* The kernel $\phi_{true}$ 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 $u_1(x) = \sin(x)\mathbf{1}_{[-\pi,\pi]}(x)$ and $u_2(x) = \sin(2x)\mathbf{1}_{[-\pi,\pi]}(x)$. When generating the data for learning, the integral $R_\phi[u_k] = f_k$ 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 $\{u_k, f_k\}_{k=1}^N = \{u_k(x_j), f_k(x_j) : x_j \in [-40, 40], j = 1, \dots, J\}_{k=1}^N$, where $x_j$ is a point on the uniform grid with mesh size $\Delta x$. For the nonlinear operator, to avoid the inverse problem being ill-defined, we introduce add an additional pair of data $(u_3, f_3)$ with $u_3(x) = x\mathbf{1}_{[-\pi,\pi]}(x)$ (see Section 5.3 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 smooth 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 $\sigma$, and the noise-to-signal-ratio is the ratio between $\sigma$ and the average $L^2$ norm of $f_k$.

**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 $\rho$. When the data mesh refines, the size of this set increases. In

Table 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 closer to 1 than when the two dis-matches. The rates are the average of the mean rates for $\text{nsr} \in \{0.1, 0.5, 1, 2\}$ in the right columns of Figure 1-3. We do not report the rate for the $l^2$ and $L^2$ regularizers because they do not have a consistent rate.

terms of the algorithm in Section 4.1, 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, the true kernels are not in this hypothesis space. Furthermore, this hypothesis space has the largest dimension for the basis matrix $B_n$ in (4.1) being non-singular, and there is no need to select an 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 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 findings are as follows.

- The SIDA-RKHS regularizer's estimator is the most accurate in most cases. However, it occasionally happens that the $l^2$ or $L^2$ regularizer performs better because of a suboptimal regularization parameter $\lambda_0$, which depends on multiple factors, ranging from the operator, numerical error, noise and treatment of the singular or ill-conditioned normal matrix, even though SIDA-RKHS regularizer is the most robust (see Figure 4 in appendix). Thus, in addition to accuracy of the estimator, it is important to also compare the convergence rates.
- 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 higher and closer to 1, as shown in Table 2.

### 5.2. Linear integral operators

We consider first the integral operator with kernel $\phi$:

$$R_\phi[u](x) = \int_\Omega \phi(|y-x|)u(y)dy = f(x).$$

After a change of variables in the integral, it is the operator $R_\phi$ in (2.2) 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 (see e.g., Evans (2010); Gin et al. (2021)) and convolution kernels in image processing in Owhadi and Yoo (2019), to name just a few.

For this operator, the exploration measure $\rho$ (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 $\rho$, 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.

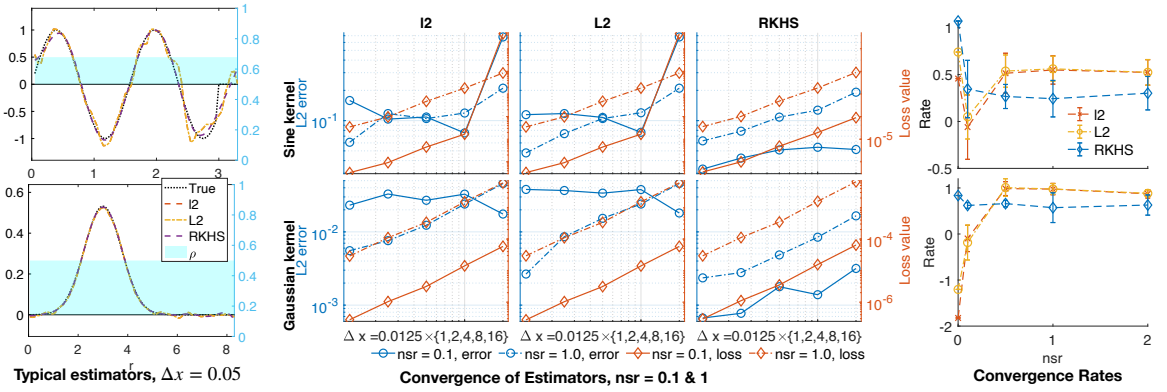


Figure 1: Linear integral operators with the sine kernel (top row) and Gaussian kernel (bottom row). Left column: typical estimators by the three regularizers, in comparison of the true kernel, superimposed with the exploration measure $\rho$ (in cyan color), when $\Delta x = 0.05$ and noise-to-signal-ratio $\mathrm{nsr} = 1$. Middle 3-columns: convergence of estimators as the mesh-size $\Delta x$ 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 $\mathrm{nsr} \in \{0, 0.1, 0.5, 1, 2\}$). Only the SIDA-RKHS regularizer's estimator consistently converges for all levels of noise, and its estimators are mostly more accurate than the other two regularizers'.

Figure 1 shows the results. The left column shows the typical estimators by the three regularizers, in comparison of the true kernel, when $\Delta x = 0.05$ and noise-to-signal-ratio $\mathrm{nsr} = 1$. The exploration measure $\rho$ (in light cyan color) is uniform for each kernel, and its support, estimated from the difference between the supports of $g[u_k]$ and $f_k$, is slightly larger than the support of the true kernel. All three regularizers lead to accurate estimators. The RKHS regularizer's estimators are the closest to the true kernel and this is further verified in the middle 3-column panel with $\Delta x = 0.05$ add $\mathrm{nsr} = 1$: for the 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 $10^{-2.5}$ while the other two regularizers' error are about $10^{-2}$.

The middle 3-column panel shows the convergence of the estimator's $L^2(\rho)$ error as the data mesh refines when $\mathrm{nsr} = 0.1$ and $\mathrm{nsr} = 1$, superimposed with the corresponding values of the loss function. When $\mathrm{nsr} = 1$, 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 $\mathrm{nsr} = 0.1$, 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 $\mathrm{nsr} = 1$ with $\Delta x \leqslant 0.025$, i.e., when the numerical error is small. Thus, these results demonstrates 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 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 closer to 1 for the smooth Gaussian kernel (which matches the continuity of data) than the rates for the discontinuous truncated sine kernel when the data are noisy. The rates are close to 1 when the data are noiseless. On the other hand, the $l^2$ and $L^2$ regularizers fails to have consistent rates when the noise level reduces. In particular, for the sine kernel, they present deceivingly higher rates than the RKHS regularizer when $\mathrm{nsr} \in \{0.5, 1, 2\}$, 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 $\Delta x$ decreases.

In short, the RKHS regularizer leads to estimators that converge consistently, at lower rates for the discontinuous sine kernel (which is discontinuous, different from the dta) 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.

### 5.3. Nonlinear operators

Next we consider the nonlinear operator $R_\phi$ with $g[u](x, y) = \partial_x[u(x + y)u(x)]$:

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

Such nonlinear operators arise in the mean-field equations of interaction particles (see e.g., Jabin and Wang (2017); Motsch and Tadmor (2014); Lu et al. (2021); Lang and Lu (2022)), and the function $\phi$ 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 $(u_3, f_3)$ 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 $f(x) = [u * \phi(|\cdot|)u]'(x)$ 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 2 shows the results. The left column shows that the exploration measure $\rho$ is non-uniform due to the nonlinear function $g[u_k]$, and 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 outperforms the other two regularizers, and they are near smooth and are close to the true kernels. The $l^2$ and $L^2$ regularizers have 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 limited data exploration. The superior performance of RKHS regularizer is further verified in the middle 3-column panel with $\Delta x = 0.05$ add $\mathrm{nsr} = 1$: its errors are much smaller than those of the other two regularizers.

The middle 3-column panel 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 shows that the RKHS regularizer has consistent rates of convergence for all levels of noises, with all rates close to 1 for the sine kernel, and slightly above 0.5 for the Gaussian kernel. In comparison, the other two regularizers' rates decreases as the noise level increases, dropping close to zero when the noise level is $\mathrm{nsr} = 2$.

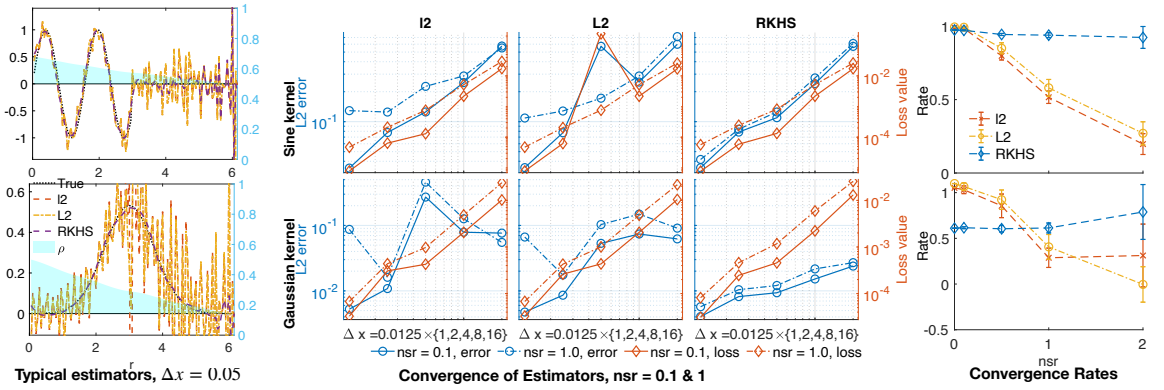


Figure 2: Nonlinear operators, in the same setting as in Figure 1. The SIDA-RKHS regularizer's estimators are significantly more accurate than those of the $l^2$ and $L^2$ regularizers in the left column. The middle 3-column panel shows that the SIDA-RKHS regularizer leads to consistently converging 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 $\mathrm{nsr} = 1$. 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.

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, different from the data). The $l^2$ and $L^2$ regularizers have convergent estimators, but the rate of convergence drops when the noise level increases.

### 5.4. Nonlocal operators

At last, we consider nonlocal operators $R_\phi$ with $g[u](x, y) = u(x + y) - u(x)$:

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

Such nonlocal operators arise in nonlocal and fractional diffusions (see e.g., Du et al. (2012); Applebaum (2009); Bucur and Valdinoci (2016)) and they have been used to construct homogenized models for peridynamic in You et al. (2022, 2020); Lu et al. (2022).

Figure 3 shows the results. The left column shows typical estimators. The exploration measure $\rho$ shrinks to zero near the origin due to the difference $g[u] = u(y) - u(x)$ and the continuity of $u$. All three regularizers lead to accurate estimators, and the RKHS estimator is the most accurate.

In the middle 3-column panel, we observe again that the RKHS regularizer leads to estimators remain converging as data mesh refines for both noise levels, even though the errors decay slower than the loss function. On the other hand, the $l^2$ and $L^2$ regularizers have inconsistent error decay: the errors decreasing monotonically when $\mathrm{nsr} = 1$, but the error lines oscillate when $\mathrm{nsr} = 0.1$ for the sine kernel, and for the Gaussian kernel, they present deceiving rates larger than the decay of the loss function due their large errors when $\Delta x$ is large.

The right column 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

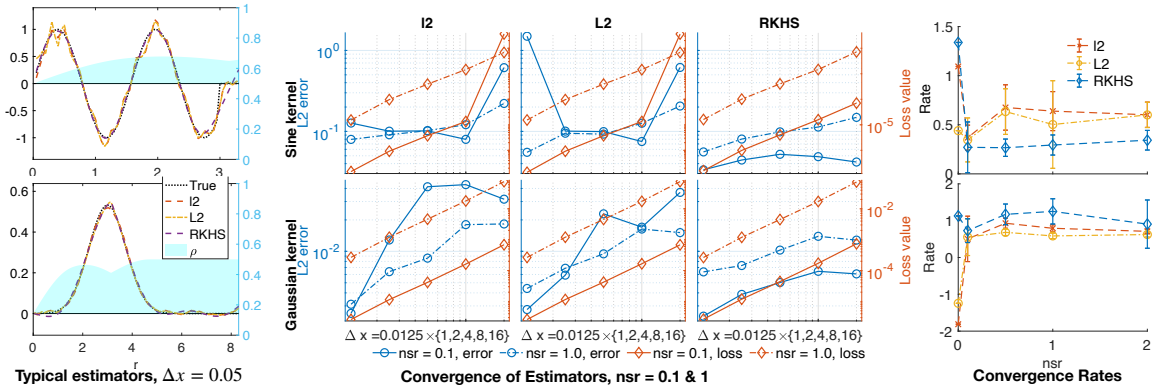


Figure 3: Nonlocal operators, in the same setting as in Figure 1. Overall, the SIDA-RKHS estimators have the smallest error mostly, and it is the only one with consistent rates for all levels of noise.

regularizer are about 0.29 for the truncated sine kernel (which has a jump discontinuity) and about 1 for the Gaussian kernel (which is continuous). Meanwhile, the rates for the $l^2$ and $L^2$-regularizers are about 0.65 for the sine kernel, and about 0.8 for the Gaussian kernel. We note again that they can have deceivingly better rates than the RKHS regularizer's while their errors are larger. Moreover, when the data is noiseless, RKHS regularizer has rates close to 1 for both kernels, while the other two regularizers rates are not consistent.

## 6. Discussion and future work

We have proposed a data adaptive RKHS Tikhonov regularization (DARTR) method for the non-parametric learning of kernel functions in operators. The DARTR method regularizes the least squares regression by the norm of a system intrinsic and data adaptive (SIDA) RKHS. It constraints the learning to take place in the function space of identifiability, in which the inverse problem is well-defined but ill-posed.

Numerical tests on synthetic datasets suggest that DARTR has the following advantages: (1) it is naturally adaptive to both data and the operator; (2) it leads to estimators converging at a consistent rate when the data mesh refines, robust to numerical error and noise.

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 are in short of a convergence analysis of the regularized estimators due to the numerical errors in the normal matrix.
2. Multivariate kernel functions. When the kernel is a multivariate function, sparse-grid representation or sparse basis functions (sparse polynomials) become necessary. A related issue is to select the optimal dimension of the hypothesis space.
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 (Zellner and Siow (1980); Bayarri et al. (2012)) that uses $\overline{A}_n^{-1}$ as prior covariance, because we have $B_{rkhs} = \overline{A}_n^{-1}$ when the basis functions are orthonormal in $L^2(\rho)$.
4. The DARTR method is applicable to general linear inverse problems with a quadratic loss functional. It is particularly useful when the data depends on the unknown function non-locally.

## Acknowledgments

The authors thank the anonymous reviewers for thoughtful comments and thank Prof. Yue Yu and Prof. Mauro Maggioni for helpful discussions. FL is grateful for the supports from NSF-DMS-1913243, FA9550-21-1-0317 and the Catalyst Award at Johns Hopkins University.

## References

David Applebaum. *Lévy processes and stochastic calculus*. Cambridge university press, 2009.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68 (3):337–404, 1950.

Frank Bauer, Sergei Pereverzev, and Lorenzo Rosasco. On regularization algorithms in learning theory. *Journal of complexity*, 23(1):52–72, 2007.

Maria J Bayarri, James O Berger, Anabel Forte, and Gonzalo García-Donato. Criteria for bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3), Jun 2012. ISSN 0090-5364.

Christian Berg, Jens Peter Reus Christensen, and Paul Ressel. *Harmonic analysis on semigroups: theory of positive definite and related functions*, volume 100. New York: Springer, 1984.

Claudia Bucur and Enrico Valdinoci. *Nonlocal Diffusion and Applications*, volume 20 of *Lecture Notes of the Unione Matematica Italiana*. Springer International Publishing, Cham, 2016. ISBN 978-3-319-28738-6 978-3-319-28739-3. doi: 10.1007/978-3-319-28739-3.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Felipe Cucker and Ding Xuan Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.

Qiang Du, Max Gunzburger, R. B. Lehoucq, and Kun Zhou. Analysis and Approximation of Nonlocal Diffusion Problems with Volume Constraints. *SIAM Rev.*, 54(4):667–696, 2012. ISSN 0036-1445, 1095-7200. doi: 10.1137/110833294.

Lawrence C Evans. *Partial differential equations*, volume 19. American Mathematical Soc., 2010.

Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*, volume 76. Springer, 2006.

Silvia Gazzola, Per Christian Hansen, and James G Nagy. Ir tools: a matlab package of iterative regularization methods and large-scale test problems. *Numerical Algorithms*, 81(3):773–811, 2019.

Craig R Gin, Daniel E Shea, Steven L Brunton, and J Nathan Kutz. Deepgreen: deep learning of green's functions for nonlinear boundary value problems. *Scientific reports*, 11(1):1–14, 2021.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

Per Christian Hansen. REGULARIZATION TOOLS: A Matlab package for analysis and solution of discrete ill-posed problems. *Numer Algor*, 6(1):1–35, 1994. ISSN 1017-1398, 1572-9265. doi: 10.1007/BF02149761.

Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*. SIAM, 1998.

Per Christian Hansen. The L-curve and its use in the numerical treatment of inverse problems. In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

Tailen Hsing and Randall Eubank. *Theoretical foundations of functional data analysis, with an introduction to linear operators*, volume 997. John Wiley & Sons, 2015.

Pierre-Emmanuel Jabin and Zhenfu Wang. Mean field limit for stochastic particle systems. In *Active Particles, Volume 1*, pages 379–402. Springer, 2017.

Hachem Kadri, Emmanuel Duflos, Philippe Preux, Stéphane Canu, Alain Rakotomamonjy, and Julien Audiffren. Operator-valued kernels for learning from functional response data. *Journal of Machine Learning Research*, 17(20):1–54, 2016.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces. *arXiv preprint arXiv:2108.08481*, 2021.

Quanjun Lang and Fei Lu. Identifiability of interaction kernels in mean-field equations of interacting particles. *arXiv preprint arXiv:2106.05565*, 2021.

Quanjun Lang and Fei Lu. Learning interaction kernels in mean-field equations of first-order systems of interacting particles. *SIAM Journal on Scientific Computing*, 44(1):A260–A285, 2022.

Zhongyang Li, Fei Lu, Mauro Maggioni, Sui Tang, and Cheng Zhang. On the identifiability of interaction functions in systems of interacting particles. *Stochastic Processes and their Applications*, 132:135–163, 2021.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Graph kernel network for partial differential equations. *arXiv preprint arXiv:2003.03485*, 2020.

Chensen Lin, Zhen Li, Lu Lu, Shengze Cai, Martin Maxey, and George Em Karniadakis. Operator learning for predicting multiscale bubble growth dynamics. *The Journal of Chemical Physics*, 154(10):104118, 2021.

Fei Lu, Mauro Maggioni, and Sui Tang. Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. *Foundations of Computational Mathematics*, pages 1–55, 2021.

Fei Lu, Qingci An, and Yue Yu. Nonparametric learning of kernels in nonlocal operators. *arXiv preprint arXiv2205.11006*, 2022.

Tom Lyche, Carla Manni, and Hendrik Speleers. *Foundations of Spline Theory: B-Splines, Spline Approximation, and Hierarchical Refinement*, volume 2219, pages 1–76. Springer International Publishing, Cham, 2018. ISBN 978-3-319-94910-9 978-3-319-94911-6. doi: 10.1007/978-3-319-94911-6_1.

Sebastien Motsch and Eitan Tadmor. Heterophilious Dynamics Enhances Consensus. *SIAM Rev*, 56(4):577 – 621, 2014.

Houman Owhadi and Gene Ryan Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.

Les Piegl and Wayne Tiller. *The NURBS Book*. Monographs in Visual Communication. Springer Berlin Heidelberg, Berlin, Heidelberg, 1997. ISBN 978-3-540-61545-3 978-3-642-59223-2. doi: 10.1007/978-3-642-59223-2.

J.O. Ramsay and B.W. Silverman. *Functional data analysis*. Springer-Verlag New York, 2 edition, 2005. ISBN ISBN : 978-0-387-40080-8.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Lei Shi, Yun-Long Feng, and Ding-Xuan Zhou. Concentration estimates for learning with $l^1$-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2):286–302, 2011.

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. ISSN 00359246. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Andrei Nikolajevits Tihonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math.*, 4:1035–1038, 1963.

Hongyan Wang. *Analysis of statistical learning algorithms in data dependent function spaces*. PhD thesis, City University of Hong Kong, 2009.

Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and Its Application*, 3:257–295, 2016.

Huaiqian You, Yue Yu, Stewart Silling, and Marta D'Elia. Data-driven learning of nonlocal models: From high-fidelity simulations to constitutive laws. *ArXiv201204157 Cs Math*, 2020.

Huaiqian You, Yue Yu, Nathaniel Trask, Mamikon Gulian, and Marta D'Elia. Data-driven learning of nonlocal physics from high-fidelity synthetic data. *Computer Methods in Applied Mechanics and Engineering*, 374:113553, 2021. ISSN 00457825. doi: 10.1016/j.cma.2020.113553.

Huaiqian You, Yue Yu, Stewart Silling, and Marta D'Elia. A data-driven peridynamic continuum model for upscaling molecular dynamics. *Computer Methods in Applied Mechanics and Engineering*, 389:114400, 2022.

Arnold Zellner and Aloysius Siow. Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, 31:585–603, 1980.

## Appendix A. Proofs

**Proof** [Proof of Lemma 3.2] Recall that a bi-variate function $\overline{G}$ is positive semi-definite if for any $(c_1, \ldots, c_m) \in \mathbb{R}^m$ and any $\{r_j\}_{j=1}^m \subset \mathbb{R}^d$, the sum $\sum_{i=1}^m \sum_{j=1}^m c_i c_j \overline{G}(r_i, r_j) \geqslant 0$. (see e.g. Berg et al. (1984); Cucker and Zhou (2007); Li et al. (2021)). Using (3.4) and (3.5), we have

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

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 and the definition of $\rho$ (see also in Lang and Lu (2021)). Also, since $\overline{G}$ is positive semi-definite, so is $\mathcal{L}_{\overline{G}}$. The equation (3.7) follows from (3.3). ■

**Proof** [Proof of Lemma 3.3] Part (a) is a standard operator characterization of the RKHS $H_G$ (see e.g., (Cucker and Zhou, 2007, Section 4.4)).

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 implies 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). ■

**Proof** [Proof of Theorem 3.6] From (3.7), we can write the loss functional in (3.1) as

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

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

For Part (b), first note that for any $\phi_N^f \in \mathcal{L}_{\overline{G}}(L^2(\rho))$, the estimator $\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1}\phi_N^f$ 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 data is continuous noiseless and the true kernel is $\phi_{true}$, i.e. $R_{\phi_{true}}[u_k] = f_k$, by (3.7) and the definition of the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ in (2.5), we have

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

for any $\psi \in L^2(\rho)$. Thus, $\phi_N^f = \mathcal{L}_{\overline{G}}\phi_{true}$ and $\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1}\phi_N^f = \phi_{true}$. That is, $\phi_{true} \in H$ is the unique minimizer of the loss functional $\mathcal{E}$. 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 $\phi^0$ such that $\mathcal{L}_{\overline{G}}\phi^0 = 0$. Thus, $H$ is the largest such function space, and we conclude that $H$ is the FSOI.

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

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

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

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

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

**Proof** [Proof of Theorem 4.1] Let $\psi_k = \sum_{j=1}^n V_{jk}\phi_j$ with $V^\top B_n V = I_n$. Then, $\psi_k$ is an eigenfunction of $\mathcal{L}_{\overline{G}}$ with eigenvalue $\lambda_k$ if and only if for each $i$,

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

where the last equality follows from the definition of $\overline{A}_n$ in (4.1). Meanwhile, by the definition of $B_n$ we have $\langle \phi_i, \lambda_k \psi_k \rangle_{L^2(\rho)} = \sum_{j=1}^n B_n(i,j)\lambda_k V_{jk}$ for each $i$. Then, Equation (4.2) follows. ■

## Appendix B. Algorithm details

### B.1. Detailed nonparametric learning algorithm

We consider only discrete data $\{u_k(x_j), f_k(x_j)\}_{k=1}^N$ in 1-dimensional and at equidistant mesh points $\{x_j = j\Delta x\}_{j=0}^J$. The extension to multi-dimensional 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 repeatedly used for all hypothesis spaces. This step can reduce the computational cost in orders of magnitude when the data is large with thousands of pairs $(u_k, f_k)$ with fine mesh.

Let $R_0$ be the diameter of the set $\Omega$. The discrete data set $\{u_k(x_j), f_k(x_j)\}_{k=1}^N$ explores only the variable $r$ of $\phi$ in the set $\mathcal{R}_N^J = \{r_{ijk} = |y_i| \leqslant R_0 : g[u_k](x_i, y_j) \neq 0 \text{ for some } i, j, k\}$, the set of all values explored by data with repetition. A discrete approximation of the exploration measure $\rho$ 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 $\rho_N^J$ uses only the information from $u_k$ and it does not reflect the information about the kernel in $f_k$. To estimate the support of the kernel, we extract the additional information from $\{f_k\}$ as follows. We set the data-adaptive support of the kernel to be $[0, R]$ with $R$ defined by

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

where $(L_i^u, R_i^u)$ and $(L_i^f, R_i^f)$ are the lower and upper bounds of the supports $g[u_k](x, y)$ and $\mathrm{supp}(f_k)$ respectively, and $R_\rho$ is the maximum of the support of $\rho_N^J$. 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 $f_k$, which extracts the additional information in the data $\{f_k\}$. 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.

The estimated support of the kernel is the region explored by data. Outside of the region, the data provides little information about the kernel. Thus, we focus on learning the kernel in this region and set the local basis functions to be supported in it. Accordingly, we constrain the exploration measure to be supported in $[0, R]$, and for simplicity of notation, we still denote it by $\rho_N^J$.

**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 $(u_k, f_k)$. 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 $r, s \in \mathrm{supp}(\rho)$. Similarly, for a basis function $\phi_i$, to compute $\overline{b}(i)$ in (2.4), which can be re-written as

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

we only need the function $g_N^f$ defined by

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

Let $r_k = k\Delta x$ for $k = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor$, which are the mesh points of $\phi$ explored by the data. Then, all the regression data we need in the original data (2.1) are

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

where $G$, $g_N^f$ and $\rho_N^J$ 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 $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$ 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 e.g., Chapter 2 of Piegl and Tiller (1997) and Lyche et al. (2018)). 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

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

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 $B_n = \mathrm{Diag}(\rho_N^J)$. 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 $\rho_N^J$.

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 $\bar{b}_n$ and the basis matrix $B_n$ as

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

The triplet $(\overline{A}_n, \bar{b}_n, B_n)$ 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. That is, our estimator is the minimizer of the regularized loss in (2.7) 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 $(\overline{A}_n, \bar{b}_n, B_n)$. It proceeds in three steps. First, we solve the generalized eigenvalue problem $\overline{A}_n V = B_n V \Lambda$, where $\Lambda$ is a diagonal matrix of the generalized eigenvalues and the matrix $V$ has columns being eigenvectors orthonormal in the sense that $V^\top B_n V = I_n$. Here these eigenvalues approximate the eigenvalue of $\mathcal{L}_{\overline{G}}$ in (3.6), and $\widehat{\psi}_k = V_{jk}\phi_j$ approximates the eigenfunctions of $\mathcal{L}_{\overline{G}}$. Then, we compute the square RKHS norm of $\phi = \sum_i c_i\phi_i$ as

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

where the inverse is taken as pseudo-inverse, particularly when $\Lambda$ has zero eigenvalues.

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

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

The regularized estimator is

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

Then, we select the hyper-parameter $\lambda$ by the L-curve method (see Section 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 $(\overline{A}_n + \lambda B_{rkhs})c_\lambda = \bar{b}_n$ can be avoided by using minimum norm least squares solution. Note that this linear equation is equivalent to $(B_{rkhs}^{-T/2}\overline{A}_n B_{rkhs}^{-1/2} + \lambda I)\tilde{c}_\lambda = B_{rkhs}^{-T/2}\bar{b}_n$ with $\tilde{c}_\lambda = B_{rkhs}^{-1/2}c_\lambda$, where $B_{rkhs}^{-T/2}$ is the transpose of the square root matrix $B_{rkhs}^{-1/2}$. Meanwhile, the square root $B_{rkhs}^{-1/2} = (V\Lambda V^\top)^{1/2}$ comes directly from (B.9). Thus, these treatments avoid the matrix inversions and lead to more robust estimators.*

We summarize the method in Algorithm 2.

**Algorithm 2** Nonparametric learning of the kernel in operator with DARTR

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

**Output:** Estimator $\widehat{\phi}$

1: Estimate the exploration measure $\rho_N^J$ 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 $(G, g_N^f)$ 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 $(G, g_N^f, \rho_N^J)$ 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 $\widehat{\phi} = \sum_{i=1}^{n^*} c_{n^*}^i \phi_i$.

### B.2. Hyper-parameter by the L-curve method

We select the parameter $\lambda$ by the L-curve method Hansen (2000); Lang and Lu (2022). 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 $\mathcal{E}(\widehat{\phi_\lambda}) = c_\lambda^\top \overline{A}_n c_\lambda - 2c_\lambda^\top \overline{b}_n - C_N^f$, and $\mathcal{R}(\phi)$ is the regularization term, for example, $\mathcal{R}(\widehat{\phi_\lambda}) = \|\widehat{\phi_\lambda}\|_{H_{\overline{G}}}^2 = c_\lambda^\top B_{rkhs} c_\lambda$. The optimal parameter is the maximizer of the curvature of $l$. In practice, we restrict $\lambda$ in the spectral range $[\lambda_{min}, \lambda_{max}]$ of the operator $\mathcal{L}_{\overline{G}}$,

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

where $\lambda_{min}$ and $\lambda_{max}$ are computed from the smallest and the largest generalized eigenvalues of $(\overline{A}_n, B_n)$. This optimal parameter $\lambda_0$ balances the loss $\mathcal{E}$ and the regularization (see Hansen (2000) for more details). Figure 4 shows a few typical L-curve plots in the selection of the parameter.

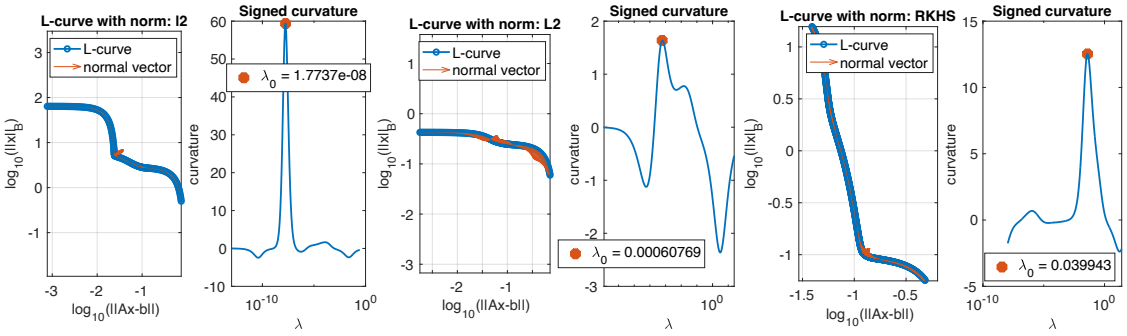


Figure 4: The L-curve selection of the optimal regularization parameter $\lambda_0$ with maximal curvature. From left to right (each with an L-curve plot and a curvature plot): $l^2$, $L^2$ and the RKHS. These optimal $\lambda_0$'s lead to the regularized estimators for the Gaussian kernel in Figure 2 (left), whose $L^2(\rho)$ errors are 0.34, 0.14, and 0.02, respectively. The RKHS-norm leads to the best shaped L-curve, and it has the most accurate estimator.