Nonparametric learning of kernels in nonlocal operators

Fei Lu[1†], Qingci An[1†] and Yue Yu[2†]

[1]Department of Mathematics, Johns Hopkins University, Baltimore, Maryland, USA .
[2]Department of Mathematics, Lehigh University, Bethlehem, Pennsylvania, USA.

Contributing authors: feilu@math.jhu.edu; qan2@jhu.edu; yuy214@lehigh.edu;
[†]These authors contributed equally to this work.

Abstract

Nonlocal operators with integral kernels have become a popular tool for designing solution maps between function spaces, due to their efficiency in representing long-range dependence and the attractive feature of being resolution-invariant. In this work, we provide a rigorous identifiability analysis and convergence study for learning kernels in nonlocal operators. It is found that kernel estimation is an ill-posed or even ill-defined inverse problem, leading to divergent estimators in the presence of modeling errors or measurement noises. To resolve this issue, we propose a nonparametric regression algorithm with a novel data adaptive RKHS Tikhonov regularization method based on the function space of identifiability. The method yields a noisy-robust convergent estimator of the kernel as the data resolution refines, on both synthetic and real-world datasets. In particular, the method successfully learns a homogenized model for stress wave propagation in a heterogeneous solid, revealing the unknown governing laws from real-world data at the microscale. Our regularization method outperforms baseline methods in robustness, generalizability, and accuracy.

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

1 Introduction

Nonlocal operators are increasingly used to represent nonlocal or long-range dependence, with numerous applications in such as nonlocal and fractional diffusion [1–9], homogenization problems [10–13], fast partial differential equations (PDE) solvers [14–17], control problems [15, 18], subsurface transport [19–23], multi-agent systems with nonlocal interaction [24–26], phase transitions [27–29], nonlocal network in machine learning [30, 31] and image processing [32–37]. Motivated by these applications, an important inverse problem emerges: learning the integral kernels of the nonlocal operators from data. Such kernel functions, with examples including the Gaussian kernel and Green's functions, are resolution-invariant and reveal the law of nonlocal interaction; thus, they are fundamental for the nonlocal operators. However, despite a long line of work on nonlocal model learning, there is limited theoretical characterization of this inverse problem, even in the linear setting. In this paper, we aim to fill the gap by studying the learning of kernels in nonlocal diffusion operators from data.

Suppose that we are given data consisting of discrete noisy observations of function pairs:

$$\mathcal{D} = \{(u_i, f_i)\}_{i=1}^N = \{(u_i(x_j), f_i(x_j)) : j = 1, \dots, J\}_{i=1}^N, \tag{1}$$

where (u_i, f_i) are pairs of real-valued continuous functions on a bounded open connected set $\Omega \subset \mathbb{R}^d$ and $\{x_j \in \Omega\}$ are spatial mesh points. The task is to learn an optimal kernel function ϕ fitting a nonlocal operator $L_\phi[u] = f$ to the data, in the form:

$$L_\phi[u](x) = \int_\Omega \phi(|y - x|)[u(y) - u(x)]dy = f(x), \forall x \in \Omega. \tag{2}$$

This operator is nonlocal in the sense that it depends on the function u nonlocally through the convolution of $u(y) - u(x)$, unlike a (local) differential operator (see Appendix C for more details). Here the data pairs can be functions, solutions to PDEs or images [10, 11, 34].

Our goal is to infer the kernel ϕ from data via nonparametric regression, so as to address the general situations that there is limited information to derive a parametric form or constraints for the kernel, which can be either smooth or singular. The regression utilizes the linear dependence of the operator on the kernel, making it possible to treat the large-size function data in a scalable fashion.

Three challenges are to be overcome. First, the function space of identifiability (FSOI) is yet to be specified properly. Otherwise, the inverse problem can be ill-defined in the sense that multiple kernels fit the data. This is fundamentally different from classical nonparametric regression that learns a function $Y = \phi(X)$ from random samples $\{(X_i, Y_i)\}$ from the joint distribution of (X, Y), for which the FSOI is $L^2(\rho)$ with ρ being the distribution of X and the optimal estimator is the conditional expectation. Second, the kernel estimator should be resolution independent and converge in a proper function space when the data resolution refines, so that it can be applied to problems and simulation

tasks with different grids or discretization methods and provides a guaranteed modeling accuracy. Third, beyond resolution invariance, the estimator should be robust to imperfect data to apply to real applications.

To overcome these challenges, we first introduce an exploration measure ρ_N, the counterpart of the measure ρ in classical regression, quantifying the exploration of the kernel's variable by data. The support of ρ_N is where there is information from data to learn the kernel. With this measure, we have an ambient function space of learning $L^2(\rho_N)$ that consists of square-integrable functions. Second, we study the identifiability of the kernel via the nonparametric regression, which leads to two findings: (1) the FSOI is data-dependent, it can be a proper subspace of $L^2(\rho_N)$, and beyond the FSOI, the inverse problem is ill-defined; (2) even in the FSOI, the inverse problem is ill-posed. Therefore, to ensure that the learning takes place inside the FSOI and to overcome the ill-posedness, we introduce a novel regularization method that uses the norm of a system-intrinsic data-adaptive reproducing kernel Hilbert space (SIDA-RKHS), whose closure is the FSOI. Finally, in experimental studies, we compare our SIDA-RKHS regularization method with two common Tikhonov/ridge regularizers that use l^2 and L^2 norms. Results on benchmark problems with synthetic data and real-world data show that only the SIDA-RKHS regularizer can consistently obtain convergent estimators for all kernels, especially when the data is noisy.

We summarize our major contributions below:

1) We establish a rigorous identifiability theory for the nonparametric learning of kernels in nonlocal operators, and specify a data-adaptive function space of identifiability (FSOI, see Lemma 3 and Theorem 1). The theory also indicates a pitfall of the nonlocal kernel learning problem: the inverse problem is ill-defined beyond the FSOI and is ill-posed in the FSOI.

2) We introduce a nonparametric regression algorithm equipped with a novel regularization method based on the SIDA-RKHS (see Section 2.3), which ensures that the learning takes place inside the FSOI and overcomes the ill-posedness to yield a convergent estimator robust to noise.

3) We validate the theory and the proposed algorithm on three benchmark problems, including various synthetic datasets and a real-world dataset where the governing law is unknown (see Section 3). Results show that the proposed algorithm provides a stable and converging estimator, while the common Tikhonov/ridge regularizers with l^2 or L^2-norm fail this task.

1.1 Related Work

Nonlocal operators: The inverse problem for nonlocal diffusion has been studied in [38, 39] from a single solution. To discover nonlocal physical laws from data, a parametric learning approach has been proposed in [11, 40, 41], where the coefficients of Bernstein polynomials are learned 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 [16, 42, 43]. 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 [10, 11, 42, 43], 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. This work tackles this issue by introducing a new regularization method based on a data-adaptive RKHS in a nonparametric learning approach.

Functional data analysis: Functional data analysis (see, e.g., [44–46] and the references therein) studies the learning of an infinite-dimensional operator from functional data. In contrast, we focus on learning a radial kernel in an operator, exploiting the low-dimensional structure of the operator, which enables us to learn the kernel (hence the operator) from limited data.

Regularization methods: Our SIDA-RKHS regularization is a Tikhonov/ridge regularization that adds a penalty term to the loss function. It differs from previous methods in the penalty term. The commonly used penalty terms include the Euclidean norm in the classical Tikhonov regularization [47, 48], the RKHS norm with an ad hoc reproducing kernel (often the Gaussian kernel) [49, 50], the total variation norm in the Rudin-Osher-Fatemi method [51], or the L^1 norm in LASSO [52]. Whereas each of these penalty terms has its specific applications, none of them take into account the FSOI, which is fundamental for learning kernels in operators. Also, our regularization method is inspired by the kernel flow method that learns hyper-parameters of the reproducing kernel [53–55], but our reproducing kernel is determined by the system and the data. Given the importance of regularization to overcome ill-posedness and overfitting, we expect our SIDA-RKHS regularization method to be applicable to a wide range of linear inverse problems and machine learning methods.

Kernel methods: The learning of kernels in nonlocal operators in this study differs from the widely used kernel methods (see, e.g.,[56–58]). The kernel methods tackle the approximation of high-dimensional functions. In contrast, our goal is to recover a resolution-invariant kernel function that reveals the low-dimensional structure of the nonlocal operator on an infinite-dimensional function space.

2 Learning theory and algorithm

2.1 Nonparametric regression with regularization

We construct an estimator by minimizing the loss functional of mean square error:

$$\mathcal{E}(\phi) = \frac{1}{N}\sum_{i=1}^{N}\int_{\Omega}|L_\phi[u_i](x) - f_i(x)|^2 dx \approx \frac{c_\Omega}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}|L_\phi[u_i](x_j) - f_i(x_j)|^2, \tag{3}$$

where the constant c_Ω depends on the mesh, e.g., $c_\Omega = \Delta x$ when $d = 1$ with uniform mesh size Δx. Hereafter, for simplicity of notation, we view (u_i, f_i) as continuous functions and write only the integral form of all elements, as long as the approximation from the discrete data is clear.

Note that the loss functional is quadratic in ϕ because the nonlocal operator is linear in ϕ. Thus, the minimizer of the loss functional is the least squares estimator (LSE), which is handy once one selects a hypothesis space $\mathcal{H}_n = \text{span}\{\phi_k\}_{k=1}^n$ with basis functions ϕ_k. Specifically, for each $\phi = \sum_{k=1}^n c_k\phi_k \in \mathcal{H}_n$, we can write the loss functional in (3) as $\mathcal{E}(c) = \mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^f$, where $C_N^f = \frac{1}{N}\sum_{i=1}^N \int |f_i(x)|^2 dx$, and the normal matrix $\overline{A}$ and vector $\overline{b}$ are given by

$$\overline{A}_n(k,l) = \langle\!\langle \phi_k, \phi_l \rangle\!\rangle, \ \overline{b}_n(k) = \frac{1}{N}\sum_{i=1}^N \int L_{\phi_k}[u_i](x) f_i(x) dx, \tag{4}$$

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

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \frac{1}{N}\sum_{i=1}^N \int_{\mathbb{R}^d} L_\phi[u_i](x) L_\psi[u_i](x) dx. \tag{5}$$

The LSE is the minimizer of $\mathcal{E}(c)$:

$$\widehat{\phi}_{\mathcal{H}_n} = \sum_{k=1}^n \widehat{c}_k \phi_k, \quad \text{where } \widehat{c} = \overline{A}_n^{-1}\overline{b}_n, \tag{6}$$

where $\overline{A}_n^{-1}$ is the inverse (or pseudo-inverse when the inverse does not exist) of $\overline{A}_n$.

However, the above least squares regression encounters a big challenge in obtaining convergent estimators for this ill-posed inverse problem (see Section 2.2). As a nonparametric method, selecting a relatively large hypothesis space is often necessary to make the model flexible enough. However, the large hypothesis space leads to a highly ill-conditioned normal matrix. As a result, the estimator in (6) oscillates violently when the data is imperfect due to either measurement noise or model error, and the estimator does not converge when the data mesh refines.

Regularization methods overcome the ill-posedness by adding a penalty term to the loss functional:

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \mathcal{R}(\phi), \tag{7}$$

where $\mathcal{R}(\phi)$ is a regularization term, and λ is a hyper-parameter controlling the contribution of the regularization term. Various penalty terms have been proposed, however, none of them take into account of the function space of identifiability, which is at the foundation of learning (see Section 2.2). Based on it, we will introduce a data-adaptive RKHS regularization method (in Section 2.3). Thus, it is different from classical regularization using an ad hoc RKHS [50, 59].

2.2 Function space of identifiability

The identifiability theory characterizes the function space of learning. There are two key elements in our identifiability theory: 1) an exploration measure, which is a probability measure that quantifies the exploration of the kernel's variable by the data, and 2) the function space of identifiability, in which the loss functional has a unique minimizer. They are described as follows.

The exploration measure. As the first key element, we introduce a novel measure on $\mathbb{R}_+$ that quantifies the exploration of the independent variable of the kernel by the data. We assume the radial kernel's support to be in an interval $[0, R_0]$. A given dataset may only explore part of this interval. More specifically, the discrete data set in (1) explores only the pairwise distances $|x_j - x_k|$ in $\mathcal{R}_N^J = \{r_{ijk} = |x_j - x_k| \leq R_0 : u_i(x_j) - u_i(x_k) \neq 0 \text{ for some } i, j, k\}$, the set of all the pairwise distances $|x_j - x_k|$ with repetition. We define an empirical measure and its continuous limit

$$\begin{aligned} \rho_N^J(dr) &= \sum_{i=1}^{N} \sum_{j,k=1}^{J} \delta_{|x_j - x_k|}(r) \frac{w_i(x_j, x_k)}{|\mathcal{R}_N^J|}, \\ \rho_N(dr) &= \sum_{i=1}^{N} \int_\Omega \int_\Omega \delta_{|x-y|}(r) \frac{w_i(x,y)}{ZN} dxdy \end{aligned} \tag{8}$$

for $r \in [0, R_0]$, where $|\mathcal{R}_N^J|$ is the cardinality of the set $\mathcal{R}_N^J$, $\delta_s(r)$ is the Kronecker delta function with value 1 when $s = r$ and with value zero otherwise, and Z is the normalizing constant. Here the weight function is $w_i(x, y) = |u_i(x) - u_i(y)|$.

The exploration measure plays an important role in the learning of the kernel. It reflects the strength of exploration to $|x - y|$ by the data $|u_i(x) - u_i(y)|$ in the loss function, and it will act as a re-weighting factor through the SIDA-RKHS regularization to be introduced in Section 2.3. Thus, we will use it to quantify the accuracy of the kernel's estimator in $L^2(\rho_N)$ (or $L^2(\rho_N^J)$ for discrete data).

Main result: function space of identifiability. We define the function space of identifiability (FSOI) as the largest linear space in which the loss functional has a unique minimizer. In other words, the variational inverse problem of finding a unique minimizer of the loss functional is well-defined in this space. In the following, we write only the continuous function space $L^2(\rho_N)$, but all the arguments apply to the discrete function space $L^2(\rho_N^J)$ in an obvious manner (see Remark 2).

Theorem 1 (Function space of identifiability) *Consider the problem of learning the kernel ϕ by minimizing the loss functional $\mathcal{E}$ in (3) with $\{u_i, f_i\}_{i=1}^N$ being continuous in a bounded domain Ω. Then, the function space of identifiability (FSOI), the largest subspace of $L^2(\rho_N)$ in which $\mathcal{E}$ has a unique minimizer, is the eigen-space of nonzero eigenvalues of $\mathcal{L}_{\overline{G}}$, an integral operator defined by*

$$\mathcal{L}_{\overline{G}}\phi(r) = \int_0^\infty \phi(s)\overline{G}(r, s)\rho_N(ds). \tag{9}$$

Here the integral kernel $\overline{G}$ *comes from data:*

$$\overline{G}(r,s) = [\rho_N'(r)\rho_N'(s)]^{-1} G(r,s), \tag{10}$$

where ρ_N' *is the density of* ρ_N *and* G *is*

$$G(r,s) = \frac{c_d^2 (rs)^{d-1}}{N} \sum_{i=1}^{N} \int_{|\eta|=1} \int_{|\xi|=1} \int [u_i(x+r\xi) - u_i(x)][u_i(x+s\eta) - u_i(x)] dx d\xi d\eta, \tag{11}$$

for $r, s \in \mathrm{supp}(\rho_N)$*, and* $G(r,s) = 0$ *otherwise, where* $c_d = 2\pi^{d/2}/(\Gamma(d/2))$ *is the area of unit sphere in* $\mathbb{R}^d$*. Furthermore, the minimizer of* $\mathcal{E}$ *is*

$$\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1} P \phi_N^f,$$

where P *is the projection to the FSOI. Here* $\phi_N^f \in L^2(\rho_N)$ *is the Riesz representation of the bounded linear functional defined by* $\langle \phi_N^f, \psi \rangle_{L^2(\rho_N)} = \frac{1}{N} \sum_{i=1}^{N} \int 2L_\psi[u_i](x) f_i(x) dx$, $\forall \psi \in L^2(\rho_N)$.

When the data is continuous and noiseless, we have $\phi_N^f = \mathcal{L}_{\overline{G}} \phi_{true}$, then, the true kernel, if it is in the FSOI, is the unique minimizer, i.e., it is identifiable by the loss functional, since $\widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1} P \phi_N^f = \phi_{true}$. When the data is discrete or noisy, the unique minimizer is an optimal estimator in the FSOI, and it is expected to converge to the true kernel when the perturbations to ϕ_N^f vanish, which happens when the data mesh refines and the noise reduces.

The proof of Theorem 1 is based on the uniqueness of zero of the Fréchet derivative of the loss functional, which becomes clear from the following lemma. Their proofs are deferred to Appendix A.

Lemma 2 *The Fréchet derivative of the loss functional* $\mathcal{E}$ *in* $L^2(\rho_N)$*, with* $\mathcal{L}_{\overline{G}}$ *defined in* (9) *and* ϕ_N^f *defined in Theorem* 1*, is* $\nabla \mathcal{E}(\phi) = 2(\mathcal{L}_{\overline{G}} \phi - \phi_N^f)$.

Remark 1 (Examples of FSOI) Here we show the FSOI in three simple cases with $N = 1$ (i.e., with a single pair of functions (u_1, f_1)) and $d = 1$. In the first two cases we consider $u_1 \equiv 1$ and $u_1(x) = x$, both of which give $f_1 \equiv 0$ for any radial kernel ϕ, revealing no information about the kernel. In either case, we have $G(r,s) \equiv 0$, which follows from (11) (or (B5) in the appendix); hence, the FSOI is a null space, detecting that the data provide no information about the kernel. In contrast, in the third case we consider $u_1(x) = x^2$, and $f(x)$ will vary with the kernel, hence revealing information about the kernel; on the other hand, we have $G(r,s) = 4r^2 s^2$, leading to a non-empty FSOI. Thus, these cases highlight the importance and meaning of the data-dependent FSOI.

System-intrinsic data-adaptive RKHS. Theorem 1 highlights two fundamental challenges: the inverse problem is well-defined only in the FSOI, and it is ill-posed in the FSOI because it involves the inverse of a compact operator $\mathcal{L}_{\overline{G}}$ (as shown in the next lemma). Fortunately, the integral kernel $\overline{G}$ defines

a reproducing kernel Hilbert space (RKHS), which provides a regularization norm to ensure the learning takes place in the FSOI and overcomes the ill-posedness. This RKHS is system intrinsic as it depends on the structure of the system of nonlocal operators, and it is data-adaptive, utilizing both the exploration measure and the data $\{u_i\}$. Thus, we call it SIDA-RKHS.

Lemma 3 (Characterization of the SIDA-RKHS) *Suppose that the data $\{u_i\}$ are continuous in Ω. Then, the following statements hold.*

(a) The integral kernel $\overline{G}$ defined in (10) is positive semi-definite.

(b) The integral operator $\mathcal{L}_{\overline{G}} : L^2(\rho_N) \to L^2(\rho_N)$ defined in (9) is compact and positive semi-definite, and we have, for any $\phi, \psi \in L^2(\rho_N)$,

$$\langle\!\langle \phi, \psi \rangle\!\rangle = \langle \mathcal{L}_{\overline{G}} \phi, \psi \rangle_{L^2(\rho_N)}. \tag{12}$$

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

(d) The eigenvalues of $\mathcal{L}_{\overline{G}}$ converges to zero, and its eigen-functions $\{\psi_k\}_k$ form a complete orthonormal basis of $L^2(\rho_N)$. For any $\phi = \sum_k c_k \psi_k$, we have

$$\langle\!\langle \phi, \phi \rangle\!\rangle = \sum_k \lambda_k c_k^2, \quad \|\phi\|^2_{L^2(\rho_N)} = \sum_k c_k^2, \quad \|\phi\|^2_{H_G} = \sum_k \lambda_k^{-1} c_k^2, \tag{13}$$

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

Remark 2 (Discrete data) When the space $L^2(\rho_N^J)$ is a discrete vector space due to discrete data, we learn the kernel on finitely many points $\{r_k\}_{k=1}^n$ explored by the data. In this case, the integral kernel G in (11) becomes a positive semi-definite matrix in $\mathbb{R}^{n\times n}$, so is $\overline{G}$ in (10). Now the operator $\mathcal{L}_{\overline{G}}$ is defined by the matrix $\overline{G}$ on the weighted vector space $\mathbb{R}^n$ and its eigenvalues is the generalized eigenvalue of $(\overline{G}, B_n)$ with B_n being a basis matrix $B_n(i,j) = \sum_{k=1}^n \phi_i(r_k)\phi_j(r_k)\rho_N^J(r_k)$, where $\{\phi_i\}$ are linearly independent basis functions of the hypothesis space $\mathcal{H}$. As a result, the FSOI is the vector space spanned by the eigenvectors with nonzero eigenvalues. Furthermore, the norm of the SIDA-RKHS H_G in (13) can be computed directly from the eigen-decomposition. This norm is better suited for regularization even when the FSOI has the same dimension as $L^2(\rho_N^J)$ (or dense in it). As data mesh refines, these vector spaces converge to the corresponding function spaces.

2.3 Algorithm: LSE with SIDA-RKHS regularization

Based on the function space of identifiability, we introduce next a nonparametric learning algorithm with SIDA-RKHS regularization. The algorithm is nonparametric in the sense that we select the best hypothesis with proper smoothness and dimension adaptive to data. Such a selection of the hypothesis space is important in recovering a resolution invariant kernel when there is

Input: The data $\{u_i, f_i\}_{i=1}^N = \{u_i(x_j), f_i(x_j)\}_{i,j=1}^{N,J}$ to construct the nonlocal model $L_\phi[u] = f$.

Output: Estimator $\widehat{\phi}$

1. Estimate the exploration measure ρ_N^J as in (8), and denote R the upper bound of its support.
2. Get regression data (see Appendix B).
3. Select a class of hypothesis spaces $\mathcal{H}_n = \text{span}\{\phi_k\}_{k=1}^n$ with n in a proper range.
4. **For** n in the range

   4a) Compute $(\overline{A}_n, \overline{b}_n, B_n)$ for $\mathcal{H}_n = \text{span}\{\phi_k\}_{k=1}^n$ with $B_n = (\langle\phi_k, \phi_l\rangle_{L^2(\rho_N^J)})_{1\le k,l\le n}$;

   4b) If the basis matrix B_n is singular, stop and remove n from the range;

   4c) Solve the generalized eigenvalue problem $\overline{A}_n V = B_n \Lambda V$, where Λ is the diagonal matrix of eigenvalues and $V^\top B_n V = I_n$;

   4d) Compute the RKHS-norm matrix $B_{rkhs} = (V\Lambda V^\top)^{-1}$;

   4e) Use the L-curve method find an optimal estimator $\widehat{\phi}_{\lambda_n^*}$.
5. Select the optimal dimension n^* (and degree if using B-spline basis) with the minimal loss value (along with other cross-validation criteria if available). Return the estimator $\widehat{\phi} = \sum_{k=1}^{n^*} c_{n^*}^k \phi_k$.

Algorithm 1: Nonparametric learning of the nonlocal kernel with SIDA-RKHS regularization

limited prior knowledge to derive a parametric kernel and requires a robust regularization method. The algorithm consists of three steps. First, we utilize the data to estimate the exploration measure and the support of the kernel. Based on them, we set a class of hypothesis spaces, with their dimensions, i.e., the number of basis functions, in a proper range moving from under-fitting to over-fitting. For each hypothesis space $\mathcal{H}_n = \text{span}\{\phi_k\}_{k=1}^n$, we compute the basis matrix $B_n = (\langle\phi_k, \phi_l\rangle_{L^2(\rho_N^J)})_{1\le k,l\le n} \in \mathbb{R}^{n\times n}$. Second, we assemble the regression matrices from data for each of these hypothesis spaces. We approximate the integrals by Riemann sum or another numerical integrator. Finally, we identify an estimator with SIDA-RKHS regularization for each of these hypothesis spaces by the L-curve method [48] and select the hypothesis space with the best fitting. We summarize the method in Algorithm 1, with its details provided in Section B.

The core innovations are the exploration measure and the regularization using the SIDA-RKHS norm. Importantly, they bring little extra computational cost. The exploration measure is available directly from data. The SIDA-RKHS norm is computed directly from the triplet $(\overline{A}_n, \overline{b}_n, B_n)$ using the generalized eigenvalue problem detailed in the algorithm.

Our SIDA-RKHS regularization uses the RKHS norm $\mathcal{R}(\phi) = c^\top B_{rkhs} c$, where B_{rkhs} is defined in (4d) in Algorithm 1. It differs from the commonly-used Tikhonov/ridge regularization using either the l^2-norm that sets $\mathcal{R}(\phi) = \sum_k c_k^2$ or the $L^2(\rho_N^J)$-norm that sets $\mathcal{R}(\phi) = c^\top B_n c$. We note that the three norms become the same when $B_n = I_n$ and all the eigenvalues of A_n are 1.

3 Tests on synthetic and real-world data

We test our method on both synthetic and real-world data in 1D examples. We compare our SIDA-RKHS regularizer on each dataset with two baseline regularizers using the l^2 and $L^2(\rho_N^J)$ norm (denoted as l2 and L2, respectively). All three regularizers use the same L-curve method to select the hyper-parameter λ as described in Appendix D.2. Importantly, they all use a projection to the FSOI (i.e., projecting the normal vector into the FSOI) to avoid an ill-defined inversion. We do not compare with other penalty norms, such as total variation or LASSO, because it is difficult to modify them to take into account the FSOI (and we leave it for future study).

In synthetic data examples, we systematically examine the method with three representative types of kernels considering both noiseless and noisy data. Since the ground-truth kernel is known, we study the convergence of estimators to the true kernel as the data mesh refines. We also apply our method to a real-world dataset for stress wave propagation in a heterogeneous bar, with the goal of constructing a homogenized model from microscale data. Since there is no ground truth, we examine the performance of estimators by studying their physical stability and capability of reproducing the wave motion on a cross-validation dataset. All datasets and codes used will be released on GitHub.

Settings for the learning algorithm. In the implementation of Algorithm 1, we use B-spline basis functions consisting of piece-wise polynomials with degree 2 so that the estimated kernel is twice differentiable (see Section D.1 for a brief introduction of B-splines). The knots of B-splines are evenly spaced on interval $[0, R]$, with one additional knot at 0 to make the first basis nonzero at $x = 0$. We select the dimension with minimal loss from a sequence of dimensions in the range $\lfloor \frac{R}{\Delta x} \rfloor \times [0.2, 1]$ as long as the basis matrix B_n is well-conditioned.

3.1 Examples with synthetic data

Numerical settings. We consider three kernels: a sine kernel, a Gaussian kernel, and a fractional Laplacian kernel (specified below). They represent bounded smooth single-scale, bounded multiscale and singular multiscale kernels, with increasing levels of challenges to learn from discrete data due to the numerical error in the approximation of the integrals. They act on the same set of functions $\{u_i\}_{i=1,2}$ with $u_1 = \sin(x)\mathbf{1}_{[-\pi,\pi]}(x)$ and $u_2(x) = \cos(x)\mathbf{1}_{[-\pi,\pi]}(x)$. In the ground-truth model, the integral $L_\phi[u_i]$ is computed by the adaptive Gauss-Kronrod quadrature method, which is much more accurate than the Riemann sum integrator that we 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 $\{u_i, f_i\}_{i=1}^N = \{u_i(x_j), f_i(x_j) : x_j \in [-40, 40], j = 1, \ldots, J\}_{i=1}^N$, where x_j is a point on the uniform grid with mesh size Δx.

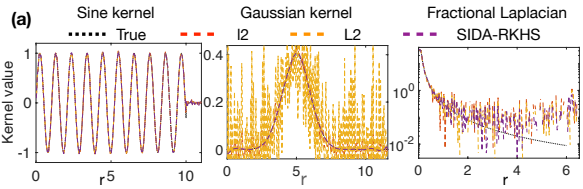
For each kernel, we consider both noiseless and noisy data with different noise levels, with a noise-to-signal ratio (nsr) taking values $\{0, 0.5, 1, 2\}$. Here the noise is added as $f_i(x_j) = L_\phi[u_i] + \eta_{i,j}$ for each i, j, where $\{\eta_{i,j}\}$ are independent and identically distributed $\mathcal{N}(0, \sigma^2)$ and the noise-to-signal-ratio is the ratio between σ and the average L^2 norm of f_i.

The three ground-truth kernels are specified as follows.

• *Sine kernel.* The sine kernel is $\phi_{true}(r) = \sin(6r)\mathbf{1}_{[0,10]}(r)$. This sine kernel represents a smooth oscillating kernel in the same class as the data u_i. The estimated support is $[0, R]$ with $R = 11.02$.

• *Gaussian kernel.* The Gaussian kernel ϕ_{true} is the Gaussian density centered at 5 with a standard deviation of 1. This kernel represents a smooth kernel. It has $R = 11.58$.

• *Fractional Laplacian kernel.* It is a truncated version of the fractional Laplacian kernel that has been widely studied in fractional and nonlocal diffusions (see, e.g., [2, 4, 6, 9]). We set $\phi_{true}(r) = c_{d,s} r^{-(d+2s)}\mathbf{1}_{[0.1,6]}(x) + 10^{d+2s}\mathbf{1}_{[0,0.1]}(x)$ with exponent $s = 0.5$ and $d = 1$, where $c_{d,s} = 4^s\pi^{-d/2}\Gamma(d/2+s)\Gamma(-s)$. It is almost singular with multiscale values and its values near the singularity are crucial to the operator. It has $R = 6.51$.



(b) Relative $L^2(\rho_N)$ error

Kernel	l2	L2	SIDA-RKHS
Sine	3.33%	**2.75%**	3.36%
Gaussian	69.6%	51.7%	**2.21%**
Fractional Laplacian	24.2%	18.7%	**17.3%**

Fig. 1 (a): Typical estimators from noisy data with noise-to-signal-ratio nsr = 1 and $\Delta x = 0.025$. (b): the relative $L^2(\rho_N^J)$ errors of these estimators. Bold numbers highlight the best method. The SIDA-RKHS regularizer consistently obtains accurate estimators in all three cases.

Performance of the regularizers. We present the typical estimators and the convergent rate of the estimator as data mesh refines, i.e., the exponent α s.t. $\|\widehat{\phi} - \phi_{true}\|_{L^2(\rho_N^J)} = \mathrm{O}((\Delta x)^\alpha)$. Figure 1 shows typical estimators for the three examples from noisy data with a noise-to-signal ratio nsr=1 and $\Delta x = 0.025$. The hypothesis space's dimension is selected by minimal loss value. All three regularizers can estimate the Sine kernel accurately and the Fractional Laplacian kernel reasonably. The SIDA-RKHS regularizer significantly outperforms the regularizers with l^2 or L^2-norm in the example of the Gaussian kernel.

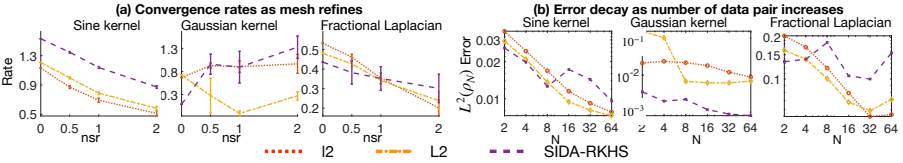


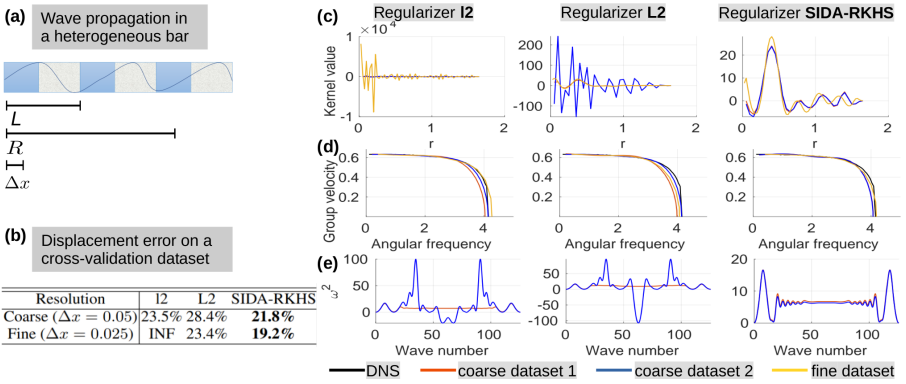
Fig. 2 (a) The means and standard deviations of the convergence rates as mesh refines in 100 independent simulations. We stress the goal is to seek an accurate estimator with a consistent rate, and the SIDA-RKHS regularizer obtains consistent rates for noisy data. We also note that the SIDA-RKHS regularizer has deceivingly lower rates for noiseless data. However, it actually has more accurate estimators (see Figure E1 in appendix). (b) Error decay as the number of data pairs increases when $nsr = 1$ and $\Delta x = 0.0125$. No convergence rate is expected here since the data are deterministic (see text).

The SIDA-RKHS regularizer's superior performance is further validated by the rates of convergence when Δx decreases, from 100 independent simulations with noises with $nsr \in \{0, 0.5, 1, 2\}$, as shown in Figure 2. It has rates generally higher than those of the other two regularizers when the data gets more noisy. Here the rates for the smooth kernels are higher than for the singular kernel, because the order of numerical error in the Riemann sum integrator is higher (see [60]).

Increasing the number of data pairs. Since the operator is linear, only linearly independent data brings new information for learning. Thus we use data $\{u_i(x)\}_{i=1}^{N} = \{\sin(ix), \cos(ix)\}_{i=1}^{N/2}$. Figure 2(b) shows that as N increases, the estimators become more accurate but without a convergence rate. Note that the data pairs do not provide independent (random) samples of ρ_N, which also varies with data. Thus, our learning problem fundamentally differs from regression for random samples, and we do not expect a convergence rate $N^{-1/2}$. An interesting future direction is to design experiments to collect informative data to enlarge the FSOI and accelerate the convergence.

In summary, the SIDA-RKHS regularizer consistently obtains accurate convergent estimators when data mesh refines for either noiseless or noisy data. On the contrary, the regularizers with l^2 norm or L^2 norm, are not robust to noise and may fail to converge, due to their negligence of the FSOI.

3.2 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%**

Fig. 3 Real-world application: wave propagation in a heterogeneous bar with the ordered microstructure of period $L = 0.4$, and the estimated support of the kernel has a bound $R = 1.65$.

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 equations, 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 in reproducing HF data in applications subject to different loading conditions with a much longer time than 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 [61], 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 unsuitable for long-term prediction because it requires the modeling of wave propagation through thousands of microstructural interfaces, making the computational cost prohibitive. To accelerate the computation, we approximate the HF model by a nonlocal model:

$$\partial_{tt} u(x,t) - L_\phi[u](x,t) = g(x,t), \text{ for } (x,t) \in \Omega \times [0,T], \tag{14}$$

where L_ϕ is a nonlocal operator in the form of (2) with a kernel ϕ being supported in $[0,R]$.

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 setting 1, 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.

Results assessment. We present the learnt estimators in Figure 3. Since there is no ground-truth kernel, we assess the performance of each estimator based on three criteria. Firstly, we report in Figure 3(b) the prediction L^2 error

of displacement on the cross validation dataset at $T = 100$. Secondly, we report in Figure 3(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(e). Its positivity indicates that the learnt nonlocal model is physically stable.

Performance of the estimators. Comparing the three estimators in Figure 3(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(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 to reproduce 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 3(b), we can see that when $\Delta x = 0.05$, all three regularizers are robust and able to reproduce the DNS simulation with 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.

3.3 Limitations and future directions

Non-radial high-dimensional kernels. When the kernel is radial, our algorithm readily applies to higher dimensions (see Appendix B). However, when the kernel is non-radial high-dimensional, the regression will face the well-known curse of dimensionality, but our identifiability theory remains valid. Thus, a future direction is to utilize kernel-regression or neural networks and further develop the SIDA-RKHS regularization.

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, numerical errors and noise, and regularization.

4 Conclusion

We have characterized the identifiability pitfall in the learning of kernels in nonlocal operators, and proposed a new regularization method to fix this issue and achieve estimator convergence. In particular, we have established

a rigorous identifiability theory for the nonparametric learning of kernels in nonlocal diffusion operators, specifying the function space of identifiability. Based on the theory, we introduced a nonparametric regression algorithm with a data-adaptive RKHS regularization method. Tests on synthetic and real-world datasets show that our algorithm consistently obtains accurate and convergent estimators, outperforming common benchmark regularizers. Our method addresses the critical estimator diverging phenomena observed in previous nonlocal operator learning methods, and the proposed framework provides a promising new direction toward overcoming the ill-posedness to achieve convergence in operator learning.

Acknowledgments. YY are supported by the National Science Foundation under award DMS 1753031, and the AFOSR grant FA9550-22-1-0197. YY would also like to like to thank Dr. Stewart Silling for his help on the DNS codes and for valuable discussions. FL is grateful for supports from NSF-1913243 and FA9550-20-1-0288. FL and QA would like to thank Quanjun Lang for helpful discussions on regularization.

Appendix A Proofs

Proof of Lemma 3 Part (a) follows directly from the definition of $\overline{G}$. Recall that a bivariate function $\overline{G}$ is positive semi-definite iff for any $(c_1, \ldots, c_m) \in \mathbb{R}^m$ and any $\{r_j\}_{j=1}^m \subset \mathbb{R}^d$, the sum $\sum_{k=1}^m \sum_{j=1}^m c_k c_j \overline{G}(r_k, r_j) \geq 0$ (see e.g. [50, 62, 63]). Then, noting that from (11) and (10) we have

$$\begin{aligned}
&\sum_{k=1}^{m}\sum_{j=1}^{m} c_k c_j \overline{G}(r_k, r_j) \\
=&\frac{1}{N}\sum_{i=1}^{N}\int_{|\eta|=1}\int_{|\xi|=1}\left[\int \sum_{k=1}^{m}\sum_{j=1}^{m} c_k c_j \frac{[u_i(x+r_k\xi)-u_i(x)][u_i(x+r_j\eta)-u_i(x)]}{\rho_N'(r_j)\rho_N'(r_k)}dx\right]d\xi d\eta \\
=&\frac{1}{N}\sum_{i=1}^{N}\int_{|\eta|=1}\int_{|\xi|=1}\left[\int \left|\sum_{k=1}^{m} c_k \frac{[u_i(x+r_k\xi)-u_i(x)]}{\rho_N'(r_k)}dx\right|^2\right]d\xi d\eta \geq 0.
\end{aligned}$$

Thus, $\overline{G}$ is positive semi-definite.

For Part (b), the operator $\mathcal{L}_{\overline{G}}$ is compact because $\overline{G} \in L^2(\rho_N \times \rho_N)$, which follows from the fact that each u_i is bounded (thus, $\overline{G}$ is also bounded). Also, since $\overline{G}$ is positive semi-definite, so is $\mathcal{L}_{\overline{G}}$. The equation (12) follows from (A2).

Part (c) is a standard operator characterization of the RKHS H_G (see e.g., [50]).

For Part (d), the eigenfunctions are orthonormal and the eigenvalues decay to zero because the operator $\mathcal{L}_{\overline{G}}$ is positive semi-definite and compact, as shown in Part (b). The first equation in (13) follows from (12), and the second equation follows from the orthogonality of the eigenfunctions. At last, if $\phi \in H_G$, by the characterization of H_G's inner product in Part (c), we have the third equation in (13). □

Proof of Lemma 2 Recall that with the bilinear form $\langle\!\langle \cdot, \cdot \rangle\!\rangle$, defined in (5), we can rewrite the loss functional as

$$\mathcal{E}(\phi) = \langle\!\langle \phi, \phi \rangle\!\rangle - \frac{1}{N}\sum_{i=1}^{N} \int 2L_\phi[u_i](x) f_i(x) dx + C_f, \tag{A1}$$

where $C_N^f = \frac{1}{N}\sum_{k=1}^{N} \int |f_i(x)|^2 dx$. Then, the derivative $\nabla\mathcal{E}(\phi)$ follows from (12) and a rewriting of the bilinear form:

$$\begin{aligned}
\langle\!\langle \phi_1, \phi_2 \rangle\!\rangle &= \frac{1}{N}\sum_{i=1}^{N} \int \left[\int\int \phi_1(|z|)[u_i(x+z) - u_i(x)]\phi_2(|y|)[u_i(x+y) - u_i(x)] dy dz \right] dx \\
&= \frac{1}{N}\sum_{i=1}^{N} \int\int \phi_1(|z|)\phi_2(|y|) \left[\int [u_i(x+z) - u_i(x)][u_i(x+y) - u_i(x)] dx \right] dy dz \\
&= \int_0^\infty \int_0^\infty \phi_1(r)\phi_2(s) G(r,s) dr ds = \int_0^\infty \int_0^\infty \phi_1(r)\phi_2(s)\overline{G}(r,s)\rho_N(dr)\rho_N(ds),
\end{aligned} \tag{A2}$$

with G and $\overline{G}$ given in (11) and (10), where the last equality is a re-weighting by ρ_N. □

Proof of Theorem 1 By Lemma 2, the Fréchet derivative of the loss functional is $\nabla\mathcal{E}(\phi) = 2(\mathcal{L}_{\overline{G}}\phi - \phi_N^f)$. Thus, the loss functional has a unique minimizer only in the function space where $\nabla\mathcal{E}(\phi)$ has a unique zero, that is, the operator $\mathcal{L}_{\overline{G}}$ has an inversion. The largest such a space is the eigenspace expanded by all eigenfunctions with non-zero eigenvalues of $\mathcal{L}_{\overline{G}}$. Furthermore, projecting ϕ_N^f to this sapce, we have the the minimizer $\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1} P\phi_N^f$ as given in the theorem. □

Appendix B Algorithm: nonparametric regression with SIDA-RKHS regularization

In this section we provide detailed description of the algorithm proposed in Section 2.3.

Our algorithm consists of three steps. First, we utilize the data to estimate the exploration measure and the support of the kernel. Based on them, we set a class of hypothesis spaces, with their dimensions i.e., the number of basis functions, in a proper range moving from under-fitting to over-fitting. Second, we assemble the regression matrices vectors from data for each of the hypothesis spaces. Finally, we identify the estimators with SIDA-RKHS regularization for these hypothesis spaces and select the one with the best fitting.

To start, we assume that the discrete data $\{u_i(x_j), f_i(x_j)\}_{i=1}^N$ comes with equidistant mesh points $\{x_j = j\Delta x\}_{j=0}^J$. For simplicity, we consider only the 1D case, and the extension to multi-dimensional cases is straightforward. We note that the current problem setting assumes data on mesh points, thus the data size increases exponentially as the dimension increases, which is the well-known curse-of-dimensionality. To overcome this curse-of-dimensionality, one

can consider other settings with mesh-free representation of data by random samples and a loss functional based on expectations (see, e.g. [64]), and this is beyond the scope of the current study.

Step 1: Set a class of hypothesis spaces.

We set a class of data-adaptive hypothesis spaces with their dimensions set to range from under-fitting to over-fitting. The key is the exploration measure and the support of the kernel estimated data. The exploration measure ρ_N^J is computed from data as in (8), which uses only the information from u_i. To estimate the support of the kernel, we extract the additional information from $\{f_i\}$ as follows. We set the data-adaptive support of the kernel to be $[0, R]$ with R defined by

$$R = 1.1 \min\{R_\rho, \max\{|L_i^f - L_i^u|, |R_i^f - R_i^u|\}_{i=1}^N\}, \tag{B3}$$

where (L_i^u, R_i^u) and (L_i^f, R_i^f) are the lower and upper bounds of the supports $\mathrm{supp}(u_i)$ and $\mathrm{supp}(f_i)$ 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_i and f_i, which extracts the additional information in the data $\{f_i\}$. 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. To avoid unbounded support in the data-based estimation in (B3), in numerical experiments we set a threshold to be 10^{-8} when estimating supports of u_i, f_i and ρ_N^J. This truncation narrows the interaction range.

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. Furthermore, we constrain the exploration measure to be supported in $[0, R]$. For simplicity of notation, we still denote it by ρ_N^J or ρ_N.

With the exploration measure and the support of the kernel, we select a class of basis functions $\{\phi_k\}_{k=1}^n$ and a range of n for the hypothesis space $\mathcal{H}_n = \mathrm{span}\{\phi_k\}_{k=1}^n$. The basis function can be either global basis functions such as Bernstein polynomials as those used in [10, 41] and trigonometric functions, or local basis functions such B-spline polynomials (see Appendix D.1 for a brief introduction). We focus on local basis functions because they are more flexible to adaptive to local structure of the kernel. 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

$$B_n = (\langle \phi_k, \phi_l \rangle_{L^2(\rho_N^J)})_{1 \le k,l \le n} \in \mathbb{R}^{n \times n} \tag{B4}$$

is non-singular. 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 non-singular while covering a wide range of dimensions. For example, when we

use piecewise constant basis, we can set $n = \lfloor \frac{R}{\Delta x} \rfloor$, 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 ρ_N^J.

Step 2: Assemble regression matrices and vectors.

We assemble the regression matrix $\overline{A}_n$ and vector $\overline{b}_n$, as defined in (4), for each hypothesis spaces $\mathcal{H}_n = \mathrm{span}\{\phi_k\}_{k=1}^n$. Together with the basis matrix B_n in (B4), the triplet $(\overline{A}_n, \overline{b}_n, B_n)$ is all we need for regression with SIDA-RKHS regularization in the next step.

To avoid repeated reading of data, we extract the regression data that can be used for all hypothesis spaces by utilizing the regression structure, which requires reading the data only once. Note that to compute $\overline{A}_n(k,k') = \langle\!\langle \phi_k, \phi_{k'} \rangle\!\rangle$ for any pair of basis functions, with the bilinear form defined in (A2), we only need G defined in (11). We note that when $d = 1$, the integral $\int_{|\eta|=1} g(\eta) d\eta = g(\eta) + g(-\eta)$, therefore, denoting $\Delta U_i(x,r) = u_i(x+r) + u_i(x-r) - 2u_i(x)$, we have

$$G(r,s) = \frac{1}{N} \sum_{i=1}^{N} \int \Delta U_i(x,r) \Delta U_i(x,s) dx \tag{B5}$$

for $r, s \in \mathrm{supp}(\rho_N)$. Similarly, for a basis function ϕ_i, to compute $\overline{b}(i)$ in (4), which can be re-written as $\overline{b}_n(k) = \frac{1}{N} \sum_{i=1}^N \int L_{\phi_k}[u_i](x) f_i(x) dx = \int_0^R \phi_k(r) g_N^f(r) dr$, we only need the function g_N^f defined by

$$g_N^f(r) = \frac{1}{N} \sum_{i=1}^{N} \int_\Omega \int_{|\xi|=1} [u_i(x + r\xi) - u_i(x)] f_i(x) d\xi \, dx. \tag{B6}$$

Let $r_l = l\Delta x$ for $l = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor$, which are all the mesh points the data explore. Then, all the regression data we need in the original data (1) are

$$\left\{ G(r_l, r_{l'}), g_N^f(r_l), \rho_N^J(r_{l'}), \text{ with } l, l' = 1, \ldots, \lfloor \frac{R}{\Delta x} \rfloor \right\}, \tag{B7}$$

where G, g_N^f and ρ_N^J are defined respectively in (11), (B6) and (8).

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

$$\begin{aligned} \overline{A}_n(k,k') = \langle\!\langle \phi_k, \phi_{k'} \rangle\!\rangle &\approx \sum_{l,l'} \phi_k(r_l) \phi_{k'}(r_{l'}) G(r_l, r_{l'})) \Delta x^2, \\ \overline{b}_n(k) &\approx \sum_{l} \phi_k(r_l) g_N^f(r_l)) \Delta x, \\ B_n(k,k') &\approx \sum_{l} \phi_k(r_l) \phi_{k'}(r_l) \rho_N^J(r_l) \Delta x. \end{aligned} \tag{B8}$$

Step 3: Regress with SIDA-RKHS regularization.

Our SIDA-RKHS regularization method uses the norm of the SIDA-RKHS so as to ensure the learning to take space in the function space of identifiability as discussed in Section 2.2. That is, our estimator is the minimizer of the regularized loss in (7) with the regularization norm $\mathcal{R}(\phi) = \|\phi\|_{H_G}^2$ defined in (13).

Computation of the RKHS norm. 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 two steps. First, we solve the generalized eigenvalue problem $\overline{A}_n V = B_n V \Lambda$, where Λ 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 (9), 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

$$\|\phi\|_{H_G}^2 = c^\top B_{rkhs} c, \text{ with } B_{rkhs} = (V \Lambda V^\top)^{-1}, \tag{B9}$$

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

$$\widehat{\phi_\lambda} = \sum_{i=1}^{n} c_\lambda^i \phi_i, \; c_\lambda = (\overline{A}_n + \lambda B_{rkhs})^{-1} \overline{b}_n. \tag{B10}$$

We will select the hyper-parameter that balances the loss $\mathcal{E}$ and the regularization term by the widely-used L-curve method [48]. It identifies the optimal hyper-parameter as the maximizer of the curvature of the curve (see Section D.2).

Appendix C Nonlocal Operators

In this section we introduce the classical and nonlocal Laplacian (diffusion) operators relevant to this paper.

Given a scalar function $u(x) : \Omega \to \mathbb{R}$, the classical Laplacian operator is defined as $\Delta u := \nabla \cdot \nabla u$ and boundary value problems on the domain Ω related to Δ are often associated with the Sobolev space $H^1(\Omega)$. On the other hand, when incorporating long-range interactions into the model such that where every point $x \in \Omega$ is interacting with a finite neighborhood of points, a nonlocal Laplacian operator is then given by

$$L_\phi[u](x) := \int_{\hat{\Omega}} \phi(x, y)(u(y) - u(x))dy, \quad x \in \Omega,$$

where $\phi(x, y)$ is a kernel function which should be specified problem by problem, $\hat{\Omega} = \Omega \cup \Omega_I$ and

$$\Omega_I := \{y \in \mathbb{R}^d \backslash \Omega \text{ such that } \phi(x, y) \neq 0 \text{ for some } x \in \Omega\}$$

is the interaction domain of Ω. This work aims to learn the kernel function ϕ from data.

In this paper we further adopt the popular choice that the interacting neighborhood of each point $x \in \Omega$ is a Euclidean ball surrounding x, i.e., $B(x, R) := \{y \in \mathbb{R}^d : |y - x| < R\}$. Here R is the interaction radius or horizon. This fact has implications on the boundary conditions that are prescribed on a collar of thickness R outside the domain Ω, that we have the interaction domain $\Omega_I = \{y \in \mathbb{R}^d \backslash \Omega : \text{dist}(y, \partial\Omega) < R\}$. Without loss of generality, we consider homogeneous Dirichlet conditions in our examples on Ω_I, i.e. $u|_{\Omega_I} = 0$. Moreover, we focus on the radial kernel such that $\phi(x, y) := \phi(|x - y|)$, which is widely employed in nonlocal problems accounting for homogenized properties (see, e.g., [2]). However, we point out that it is actually straightforward, with more complicated notations and labor in coding, to extend the current framework to non-equidistant cases or low-dimensional non-radial cases.

Appendix D B-spline basis functions and the L-curve method

D.1 B-spline basis functions

B-spline is a class of piecewise polynomials, and is capable of representing the local information of the target function. Here we review briefly the recurrence definition and properties of the balanced B-splines, for more details we refer to the Chapter 2 of [65] and [66].

Given a non-decreasing sequence of real numbers $\{r_0, r_1, \ldots, r_m\}$ (called knots), the B-spline basis functions of degree p, denoted by $\{N_{i,p}\}_{i=0}^{m-p}$, is defined recursively as

$$\begin{aligned} N_{i,0}(r) &= \begin{cases} 1, & r_i \le r < r_{i+1}, \\ 0, & otherwise, \end{cases} \\ N_{i,p}(r) &= \frac{r - r_i}{r_{i+p} - r_i} N_{i,p-1}(r) + \frac{r_{i+p+1} - r}{r_{i+p+1} - r_{i+1}} N_{i+1,p-1}(r). \end{aligned} \tag{D11}$$

The B-spline basis has the following properties:

- 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 smoothness increases with the degree but decreases when the knot multiplicity increases;
- The basis satisfies partition unity: 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 to be a uniform partition of the support of $\overline{\rho}$, $[R_{min}, R_{max}]$,

$$R_{min} = r_0 \le r_1 \le \cdots \le r_m = R_{min}.$$

We set the basis functions of the hypothesis $\mathcal{H}$, whose dimension is $n = m - p$, to be

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

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

D.2 Hyper-parameter selection by the L-curve method

We select the parameter λ by the L-curve method [48, 64]. 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, $\mathcal{R}(\widehat{\phi_\lambda}) = \|\widehat{\phi_\lambda}\|^2_{H_{\overline{G}}} = c_\lambda^\top B_{rkhs} c_\lambda$. 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 $\mathcal{L}_{\overline{G}}$,

$$\lambda_0 = \underset{\lambda_{\min} \le \lambda \le \lambda_{\max}}{\arg\max}\ \kappa(l(\lambda)) = \underset{\lambda_{\min} \le \lambda \le \lambda_{\max}}{\arg\max}\ \frac{x'y'' - x'y''}{(x'^{\,2} + y'^{\,2})^{3/2}}, \tag{D12}$$

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 [48] 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].

Appendix E Additional numerical results for synthetic data examples

This section provides additional numerical results for the examples with synthetic data.

Figure E1 shows that the SIDA-RKHS regularizer leads to converging estimators in all three examples for both noisy and noiseless data, whereas the l^2-norm and the L^2-norm regularizers' estimators have slow convergent rates or even no convergence when the data is noisy.

We note that the performance of these regularizers depends on the optimal regularization strength λ_0, which is selected by the L-curve method introduced in Section D.2. 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 E2 for typical L-curve plots).

[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.

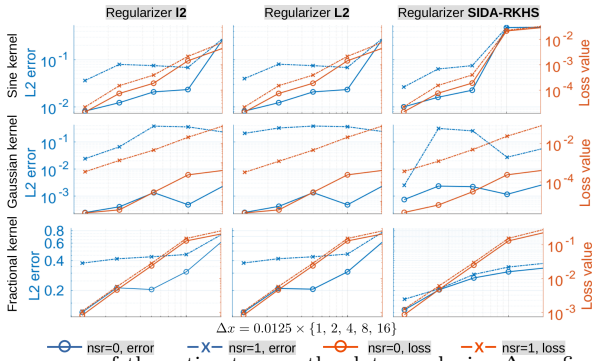


Fig. E1 Convergence of the estimators as the data mesh-size Δx refines, along with the values of the loss function. The SIDA-RKHS regularizer consistently converges for both noiseless and noisy data, with better rates (slope) than the other two regularizers for noisy data. Note that for the fractional kernel, it has a lower rate though being more accurate.

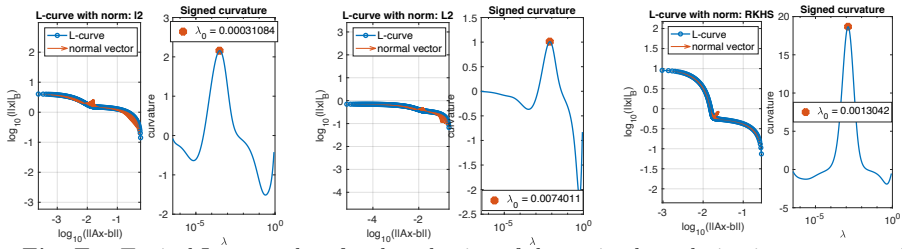


Fig. E2 Typical L-curve plots for the selection of the optimal regularization parameter λ_0 for the Gaussian kernel with $\Delta x = 0.05$ and $nsr = 1$. From left to right: the l^2, L^2 and SIDA-RKHS regularizers. All regularizers successfully select the optimal λ_0, and the SIDA-RKHS regularizer has the most well-shaped L-curve.

Appendix F Detailed Real-world Dataset Experiment Settings

In this section we provide further experiment details for the real-world dataset studied in 3.2.

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 (14). The HF-model is a classical wave equation: the displacement $\hat{u}(x,t)$ satisfies, for $(x,t) \in \Omega \times [0,T]$ with $\Omega \subset \mathbb{R}$,

$$\partial_{tt}\hat{u}(x,t) - L_{HF}[\hat{u}](x,t) = g(x,t), \tag{F13}$$

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, (F13) 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 (F13) by the direct numerical solver (DNS) introduced

in [61]. 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 $\{\hat{u}_i, g_i\}_{i=1}^N = \{\hat{u}_i(x_j, t^n), g_i(x_j, t^n) : j = 1, \ldots, J\}_{i=1,n=0}^{N,T/\Delta t}$ 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 L_ϕ that approximates the operator L_{HF} from data $\{\hat{u}, f\}$ generated by $L_{HF}[\hat{u}] = f$, where $f = \partial_{tt}\hat{u} - g$. Discretizing the time derivative in (14) with the central difference scheme, we obtain

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

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

$$\mathcal{E}(\phi) = \frac{\Delta t}{NT} \sum_{k=1}^{N} \sum_{n=1}^{T/\Delta t} \|L_\phi[\hat{u}_k^n] - f_k^n\|_{L^2(\Omega)}^2. \tag{F14}$$

F.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 $L = 0.2$, $\Delta t = 0.02$, $E_1 = 1$, $E_2 =$, $\rho = 1$, 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 G 2D MD dataset:

On this 2D dataset, the nonlocal model for the the $\mathbb{R}^2$-valued field $\mathbf{u}$ writes:

$$\begin{aligned}\mathcal{L}_\phi[\mathbf{u}](\mathbf{x}) :=& -\frac{2}{m(\delta)}\int_{B_\delta(\mathbf{x})}(\lambda-\mu)\,\phi(|\mathbf{y}-\mathbf{x}|)\,(\mathbf{y}-\mathbf{x})\,(\theta(\mathbf{x})+\theta(\mathbf{y}))\,d\mathbf{y}\\ &-\frac{16}{m(\delta)}\int_{B_\delta(\mathbf{x})}\mu\phi(|\mathbf{y}-\mathbf{x}|)\frac{(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^2}\,(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y}=\mathbf{f}(\mathbf{x}),\end{aligned} \tag{G15}$$

where scalar nonlocal dilation θ is

$$\begin{aligned}\theta(\mathbf{x}) &:= \frac{2}{m(\delta)}\int_{B_\delta(\mathbf{x})}\phi(|\mathbf{y}-\mathbf{x}|)(\mathbf{y}-\mathbf{x})\cdot(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y},\\ m(\delta) &:= \int_{B_\delta(\mathbf{0})}\phi(|\mathbf{z}|)|\mathbf{z}|^2 d\mathbf{z}.\end{aligned} \tag{G16}$$

Unknowns include: $\phi(|\mathbf{z}|)$, $\lambda = E\nu/(1-\nu^2)$, $\mu = E/(2(1+\nu))$, here E and ν are Young's modulus and Poisson ratio, respectively.

To simplify the problem, we can approximate the nonlocal dilatation, θ, with the local one, i.e., rewrite $\theta(\mathbf{x}) \approx \nabla\cdot\mathbf{u}(\mathbf{x})$. Moreover, since $m(\delta)$ is a scalar constant, we can treat $(\lambda-\mu)/m(\delta) := \alpha$ and $\mu/m(\delta) := \beta$ as trainable parameters, then the nonlocal model writes:

$$\begin{aligned}\mathcal{L}_\phi[\mathbf{u}](\mathbf{x}) :=& -2\int_{B_\delta(\mathbf{x})}\alpha\phi(|\mathbf{y}-\mathbf{x}|)\,(\mathbf{y}-\mathbf{x})\,(\nabla\cdot\mathbf{u}(\mathbf{x})+\nabla\cdot\mathbf{u}(\mathbf{y}))\,d\mathbf{y}\\ &-16\int_{B_\delta(\mathbf{x})}\beta\phi(|\mathbf{y}-\mathbf{x}|)\frac{(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^2}\,(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y}=\mathbf{f}(\mathbf{x}).\end{aligned} \tag{G17}$$

Moreover, to further make this problem a linear regression, we might consider using the Poisson ratio from literature, i.e., $\nu = -0.43$ for temperature $0K$ and $\nu = -0.42$ for temperature $300K$, so we have a fixed ratio between α and β as:

$$\alpha/\beta = (\lambda-\mu)/\mu = (3\nu-1)/(1-\nu).$$

Taking the temperature $0K$ data for example, we will have $\alpha/\beta = -1.6011$. Denoting $\tilde{\phi} := \beta\phi$, problem (G17) can then be reformulated as:

$$\begin{aligned}\mathcal{L}_{\tilde{\phi}}[\mathbf{u}](\mathbf{x}) :=& -3.2022 \int_{B_\delta(\mathbf{x})} \tilde{\phi}(|\mathbf{y}-\mathbf{x}|)\,(\mathbf{y}-\mathbf{x})\,(\nabla\cdot\mathbf{u}(\mathbf{x}) + \nabla\cdot\mathbf{u}(\mathbf{y}))\,d\mathbf{y} \\ & - 16\int_{B_\delta(\mathbf{x})} \tilde{\phi}(|\mathbf{y}-\mathbf{x}|)\frac{(\mathbf{y}-\mathbf{x})\otimes(\mathbf{y}-\mathbf{x})}{|\mathbf{y}-\mathbf{x}|^2}\,(\mathbf{u}(\mathbf{y})-\mathbf{u}(\mathbf{x}))\,d\mathbf{y} = \mathbf{f}(\mathbf{x}) \\ =& \int_{B_\delta(\mathbf{x})} \tilde{\phi}(|\mathbf{y}-\mathbf{x}|) g[\mathbf{u}](\mathbf{x},\mathbf{y}) d\mathbf{y},\end{aligned} \tag{G18}$$

where the $\mathbb{R}^2$-valued function $g[\mathbf{u}](\mathbf{x},\mathbf{y})$ is defined accordingly.

In computation: we assemble the regression matrix and vector as in Step 2 in Section B. Suppose we have the discrete observations of u_n and f_n at uniform grid $\{x_{i,j}\}_{i,j=1,\cdots,J}$ with $h = x^1_{i+1,j} - x^1_{i,j} = x^2_{i,j+1} - x^2_{i,j}$. With the uniform mesh, we have pairwise distances $\{h, \sqrt{2}h, 2h, \sqrt{5}h, 2\sqrt{2}h, 3h, \dots, \} = \{\sqrt{i^2+j^2}h, 0 \le i,j \le J\}$. Then, we count the different pairwise distances, and assemble radial G. When the number of different pairwise distances is large (e.g., more than 5000, which would lead a large non-sparse matrix), we use a histogram with a specified number of bins to control the size of G. Denote the distinct pairwise distances by $\{r_l\}_{l=1}^L$. Then, we proceed as in (B7) and (B8).

References

[1] Silling, S.A., Epton, M., Weckner, O., Xu, J., Askari, E.: Peridynamic States and Constitutive Modeling. J Elasticity **88**(2), 151–184 (2007). https://doi.org/10.1007/s10659-007-9125-1

[2] Du, Q., Gunzburger, M., Lehoucq, R.B., Zhou, K.: Analysis and Approximation of Nonlocal Diffusion Problems with Volume Constraints. SIAM Rev. **54**(4), 667–696 (2012). https://doi.org/10.1137/110833294

[3] D'Elia, M., Du, Q., Glusa, C., Gunzburger, M., Tian, X., Zhou, Z.: Numerical methods for nonlocal and fractional models. Acta Numerica **29**, 1–124 (2020). https://doi.org/10.1017/S096249292000001X

[4] Applebaum, D.: Lévy Processes and Stochastic Calculus. Cambridge university press (2009)

[5] Andreu-Vaillo, F., Mazón, J., Rossi, J., Toledo-Melero, J.J.: Nonlocal Diffusion Problems. Mathematical Surveys and Monographs, vol. 165. American Mathematical Society, Providence, Rhode Island (2010). https://doi.org/10.1090/surv/165

[6] Bucur, C., Valdinoci, E.: Nonlocal Diffusion and Applications. Lecture Notes of the Unione Matematica Italiana, vol. 20. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-28739-3

[7] Chen, Z.-Q., Zhang, X.: Heat Kernels for Non-symmetric Non-local Operators. In: Palatucci, G., Kuusi, T. (eds.) Recent Developments in Nonlocal Theory, pp. 24–51. De Gruyter Open (2017). https://doi.org/10.1515/9783110571561-003

[8] Xiong, J., Zheng, J., Zhou, X.: Unique strong solutions of lévy processes driven stochastic differential equations with discontinuous coefficients. Stochastics **91**(4), 592–612 (2019)
[9] Wang, H., Basu, T.S.: A fast finite difference method for two-dimensional space-fractional diffusion equations. SIAM Journal on Scientific Computing **34**(5), 2444–2458 (2012)
[10] You, H., Yu, Y., Trask, N., Gulian, M., D'Elia, M.: Data-driven learning of nonlocal physics from high-fidelity synthetic data. Computer Methods in Applied Mechanics and Engineering **374**, 113553 (2021). https://doi.org/10.1016/j.cma.2020.113553
[11] You, H., Yu, Y., Silling, S., D'Elia, M.: A data-driven peridynamic continuum model for upscaling molecular dynamics. Computer Methods in Applied Mechanics and Engineering **389**, 114400 (2022)
[12] Lin, C., Maxey, M., Li, Z., Karniadakis, G.E.: A seamless multiscale operator neural network for inferring bubble dynamics. Journal of Fluid Mechanics **929** (2021)
[13] Lin, C., Li, Z., Lu, L., Cai, S., Maxey, M., Karniadakis, G.E.: Operator learning for predicting multiscale bubble growth dynamics. The Journal of Chemical Physics **154**(10), 104118 (2021)
[14] Lu, L., Jin, P., Karniadakis, G.E.: Deeponet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. arXiv preprint arXiv:1910.03193 (2019)
[15] Lu, L., Jin, P., Pang, G., Zhang, Z., Karniadakis, G.E.: Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. Nature Machine Intelligence **3**(3), 218–229 (2021)
[16] Li, Z., Kovachki, N.B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A., *et al.*: Fourier neural operator for parametric partial differential equations. In: International Conference on Learning Representations (2020)
[17] Kovachki, N., Li, Z., Liu, B., Azizzadenesheli, K., Bhattacharya, K., Stuart, A., Anandkumar, A.: Neural operator: Learning maps between function spaces. arXiv preprint arXiv:2108.08481 (2021)
[18] Hwang, R., Lee, J.Y., Shin, J.Y., Hwang, H.J.: Solving pde-constrained control problems using operator learning. arXiv preprint arXiv:2111.04941 (2021)
[19] Benson, D.A., Wheatcraft, S.W., Meerschaert, M.M.: Application of a fractional advection-dispersion equation. Water Resources Research **36**(6), 1403–1412 (2000)
[20] Katiyar, A., Agrawal, S., Ouchi, H., Seleson, P., Foster, J.T., Sharma, M.M.: A general peridynamics model for multiphase transport of non-Newtonian compressible fluids in porous media. Journal of Computational Physics (2019). In press.
[21] Katiyar, A., Foster, J.T., Ouchi, H., Sharma, M.M.: A peridynamic formulation of pressure driven convective fluid transport in porous media. Journal of Computational Physics **261**, 209–229 (2014)

[22] Schumer, R., Benson, D.A., Meerschaert, M.M., Baeumer, B.: Multiscaling fractional advection-dispersion equations and their solutions. Water Resources Research **39**(1), 1022–1032 (2003)
[23] Schumer, R., Benson, D.A., Meerschaert, M.M., Wheatcraft, S.W.: Eulerian derivation of the fractional advection-dispersion equation. Journal of Contaminant Hydrology **48**, 69–88 (2001)
[24] Lu, F., Zhong, M., Tang, S., Maggioni, M.: Nonparametric inference of interaction laws in systems of agents from trajectory data. Proc. Natl. Acad. Sci. USA **116**(29), 14424–14433 (2019)
[25] Lu, F., Maggioni, M., Tang, S.: Learning interaction kernels in stochastic systems of interacting particles from multiple trajectories. Foundations of Computational Mathematics, 1–55 (2021)
[26] Lang, Q., Lu, F.: Identifiability of interaction kernels in mean-field equations of interacting particles. arXiv preprint arXiv:2106.05565 (2021)
[27] Bates, P.W., Chmaj, A.: An integrodifferential model for phase transitions: stationary solutions in higher space dimensions. Journal of Statistical Physics **95**, 1119–1139 (1999)
[28] Chen, C.K., Fife, P.C.: Nonlocal models of phase transitions in solids. Advances in Mathematical Sciences and Applications **10**(2), 821–849 (2000)
[29] Dayal, K., Bhattacharya, K.: Kinetics of phase transformations in the peridynamic formulation of continuum mechanics. Journal of the Mechanics and Physics of Solids **54**(9), 1811–1842 (2006)
[30] Wang, X., Girshick, R., Gupta, A., He, K.: Non-local Neural Networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7794–7803. IEEE, Salt Lake City, UT, USA (2018). https://doi.org/10.1109/CVPR.2018.00813
[31] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Neural Operator: Graph Kernel Network for Partial Differential Equations. ArXiv200303485 Cs Math Stat (2020) https://arxiv.org/abs/2003.03485 [cs, math, stat]
[32] Buades, A., Coll, B., Morel, J.M.: Image denoising methods: a new nonlocal principle. SIAM Review **52**, 113–147 (2010)
[33] Gilboa, G., Osher, S.: Nonlocal linear image regularization and supervised segmentation. Multiscale Modeling and Simulation **6**, 595–630 (2007)
[34] Lou, Y., Zhang, X., Osher, S., Bertozzi, A.: Image recovery via nonlocal operators. Journal of Scientific Computing **42**, 185–197 (2010)
[35] Kindermann, S., Osher, S., Jones, P.W.: Deblurring and denoising of images by nonlocal functionals. Multiscale Modeling & Simulation **4**(4), 1091–1115 (2005)
[36] Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. Multiscale Modeling & Simulation **7**(3), 1005–1028 (2009)
[37] Holler, G., Kunisch, K.: Learning nonlocal regularization operators. ArXiv200109092 Math (2020) https://arxiv.org/abs/2001.09092 [math]
[38] Jin, B., Rundell, W.: A tutorial on inverse problems for anomalous diffusion

processes. Inverse problems **31**(3), 035003 (2015)
[39] Li, Y., Lu, Y., Xu, S., Duan, J.: Extracting stochastic dynamical systems with *alpha*-stable lévy noise from data. arXiv preprint arXiv:2109.14881 (2021)
[40] Xu, X., D'Elia, M., Glusa, C., Foster, J.T.: Machine-learning of nonlocal kernels for anomalous subsurface transport from breakthrough curves. arXiv preprint arXiv:2201.11146 (2022)
[41] You, H., Yu, Y., Silling, S., D'Elia, M.: Data-driven learning of nonlocal models: From high-fidelity simulations to constitutive laws. ArXiv201204157 Cs Math (2020) https://arxiv.org/abs/2012.04157 [cs, math]
[42] Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., Anandkumar, A.: Neural operator: Graph kernel network for partial differential equations. arXiv preprint arXiv:2003.03485 (2020)
[43] You, H., Yu, Y., D'Elia, M., Gao, T., Silling, S.: Nonlocal kernel network (nkn): a stable and resolution-independent deep neural network. arXiv preprint arXiv:2201.02217 (2022)
[44] Hsing, T., Eubank, R.: Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators vol. 997. John Wiley & Sons (2015)
[45] Kadri, H., Duflos, E., Preux, P., Canu, S., Rakotomamonjy, A., Audiffren, J.: Operator-valued kernels for learning from functional response data. Journal of Machine Learning Research **17**(20), 1–54 (2016)
[46] Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice vol. 76. Springer (2006)
[47] Hansen, P.C.: REGULARIZATION TOOLS: A Matlab package for analysis and solution of discrete ill-posed problems. Numer Algor **6**(1), 1–35 (1994). https://doi.org/10.1007/BF02149761
[48] Hansen, P.C.: The L-Curve and Its Use in the Numerical Treatment of Inverse Problems, pp. 119–142. WIT Press (2000)
[49] Cucker, F., Smale, S.: On the mathematical foundations of learning. Bulletin of the American mathematical society **39**(1), 1–49 (2002)
[50] Cucker, F., Smale, S.: Emergent behavior in flocks. IEEE Transactions on Automatic Control **52**(5), 852–862 (2007). https://doi.org/10.1109/TAC.2007.895842
[51] Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena **60**(1-4), 259–268 (1992)
[52] Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 267–288 (1996). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
[53] Owhadi, H., Yoo, G.R.: Kernel Flows: From learning kernels from data into the abyss. Journal of Computational Physics **389**, 22–47 (2019). https://doi.org/10.1016/j.jcp.2019.03.040

[54] Hamzi, B., Owhadi, H.: Learning dynamical systems from data: A simple cross-validation perspective, part I: Parametric kernel flows. Physica D: Nonlinear Phenomena **421**, 132817 (2021). https://doi.org/10.1016/j.physd.2020.132817

[55] Chen, Y., Owhadi, H., Stuart, A.M.: Consistency of Empirical Bayes And Kernel Flow For Hierarchical Parameter Estimation. ArXiv200511375 Cs Math Stat (2021) https://arxiv.org/abs/2005.11375 [cs, math, stat]

[56] Le, L., Hao, J., Xie, Y., Priestley, J.: Deep kernel: learning kernel function from data using deep neural network. In: Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, pp. 1–7 (2016)

[57] Atanasov, A., Bordelon, B., Pehlevan, C.: Neural networks as kernel learners: The silent alignment effect. arXiv preprint arXiv:2111.00034 (2021)

[58] Williams, C.K.I., Rasmussen, C.E.: Gaussian Processes for Regression, 7

[59] Bauer, F., Pereverzev, S., Rosasco, L.: On regularization algorithms in learning theory. Journal of complexity **23**(1), 52–72 (2007)

[60] Fan, Y., Tian, X., Yang, X., Li, X., Webster, C., Yu, Y.: An asymptotically compatible probabilistic collocation method for randomly heterogeneous nonlocal problems. arXiv preprint arXiv:2107.01386 (2021)

[61] Silling, S.A.: Propagation of a stress pulse in a heterogeneous elastic bar. Sandia Report SAND2020-8197, Sandia National Laboratories (2020)

[62] Berg, C., Christensen, J.P.R., Ressel, P.: Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions vol. 100. New York: Springer (1984)

[63] Li, Z., Lu, F., Maggioni, M., Tang, S., Zhang, C.: On the identifiability of interaction functions in systems of interacting particles. Stochastic Processes and their Applications **132**, 135–163 (2021)

[64] Lang, Q., Lu, F.: Learning interaction kernels in mean-field equations of first-order systems of interacting particles. SIAM Journal on Scientific Computing **44**(1), 260–285 (2022)

[65] Piegl, L., Tiller, W.: The NURBS Book. Monographs in Visual Communication. Springer Berlin Heidelberg, Berlin, Heidelberg (1997). https://doi.org/10.1007/978-3-642-59223-2

[66] Lyche, T., Manni, C., Speleers, H.: Foundations of Spline Theory: B-Splines, Spline Approximation, and Hierarchical Refinement, vol. 2219, pp. 1–76. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-94911-6_1