Data-driven learning of robust nonlocal physics from high-fidelity synthetic data

Huaiqian You[a], Yue Yu[a,], Nathaniel Trask[b], Mamikon Gulian[b], Marta D'Elia[c]

[a]*Department of Mathematics, Lehigh University, Bethlehem, PA*
[b]*Center for Computing Research, Sandia National Laboratories, Albuquerque, NM*
[c]*Computational Science and Analysis, Sandia National Laboratories, Livermore, CA*

Abstract

A key challenge to nonlocal models is the analytical complexity of deriving them from first principles, and frequently their use is justified a posteriori. In this work we extract nonlocal models from data, circumventing these challenges and providing data-driven justification for the resulting model form. Extracting provably robust data-driven surrogates is a major challenge for machine learning (ML) approaches, due to nonlinearities and lack of convexity. Our scheme allows extraction of provably invertible nonlocal models whose kernels may be partially negative. To achieve this, based on established nonlocal theory, we embed in our algorithm sufficient conditions on the non-positive part of the kernel that guarantee well-posedness of the learnt operator. These conditions are imposed as inequality constraints and ensure that models are robust, even in small-data regimes. We demonstrate this workflow for a range of applications, including reproduction of manufactured nonlocal kernels; numerical homogenization of Darcy flow associated with a heterogeneous periodic microstructure; nonlocal approximation to high-order local transport phenomena; and approximation of globally supported fractional diffusion operators by truncated kernels.

Keywords: nonlocal models, data-driven learning, multiscale modeling, machine learning, operator regression, optimization, homgenization, fractional models

1. Background

In contrast to partial differential equation (PDE) models which typically govern classical continuum mechanics and standard diffusion, nonlocal models describe systems in terms of integro-differential equations (IDEs). IDEs possess several modeling advantages over

---

Email addresses: `huy316@lehigh.edu` (Huaiqian You), `yuy214@lehigh.edu` (Yue Yu), `natrask@sandia.gov` (Nathaniel Trask), `mgulian@sandia.gov` (Mamikon Gulian), `mdelia@sandia.gov` (Marta D'Elia)

PDEs: reduced regularity requirements admit simpler descriptions of fracture mechanics[1–3]; integral operators allow more natural description of long-range molecular interactions such as electrokinetic and surface tension effects [4–8]; and compactly supported integral operators introduce a modeling lengthscale δ which may be used to model subgrid microstructures [9–13]. We consider in this work compactly supported nonlocal diffusion models, extensively analyzed in the literature [14–16]. While they have been effectively used, it is in general unclear how to justify the use of a kernel for a given application. As an example, in the peridynamic theory of continuum mechanics, the choice of nonlocal kernel is often justified a posteriori. Free parameters in constitutive models, including the lengthscale δ, are often tuned to obtain agreement with experiment. In this manner, peridynamics has yielded enhanced descriptions of fracture mechanics, composite mechanics, but it is an open question how to obtain predictive nonlocal models from first principles.

Recent work presents compelling justification of nonlocal kernels' role bridging microstructure and homogenized continuum response. Weckner and Silling characterize the multiscale response of elastic media by calibrating nonlocal dispersion relations to neutron scattering data data [17]. Delgoshaie argued that nonlocal diffusion kernels naturally describe tortuous flow pathways in porous media [18], while Chung obtained a nonlocal description of flow in fractured media via a multiscale finite element framework [19]. Several authors have established conditions where nonlocal diffusion kernels admit interpretation as jump rates of stochastic processes [20–22]. Recently, Xu and Foster derived kernels describing an elastic medium with heterogeneous microstructure via homogenization[23].

Works which provide rigorous a priori justification of nonlocal kernels typically require mathematically complex and lengthy derivations, often relying on restrictive assumptions of microstructure geometry. The goal of the current work is to establish a data-driven extraction of nonlocal models from observations of high-fidelity simulations, skirting the mathematical complexity of deriving nonlocal models. We pursue an inequality-constrained least squares approach which guarantees that extracted models are well-posed. In addition to learning unknown physics from systems, our method may be used to perform numerical homogenization, extracting efficient coarse-grained nonlocal models from high-fidelity synthetic data, such as microstructure-resolved local simulations. Furthermore, our algorithm may be applied to learn computationally cheap approximations of nonlocal operators of fractional-type while preserving accuracy. The latter are characterized by an infinite lengthscale and, as such, are computationally prohibitively expensive. Our technique delivers compactly supported nonlocal models that mimic the action of a fractional Laplacian at a much cheaper cost.

An additional open question in nonlocal modeling is whether kernels should be strictly positive. In some cases, nonlocal solutions can be interpreted as probability density functions

of jump processes with jump rate given by the kernel, which, in turn, must be positive [20–22]. On the other hand, in multiscale problems, several authors have extracted sign-changing kernels [17]; see also the discussion in [24]. It is unclear however, whether multiscale physics inherently lead to sign-changing kernels or if one could derive equally descriptive positive kernels. We note also that the mathematical analysis of non-positive kernels is more involved, suggesting that strictly positive kernels may be more desirable to work with [24]. Recently, Mengesha and Du provided necessary conditions for diffusion processes governed by sign-changing kernels to be well-posed [24]. Their analysis requires that the kernel may be decomposed into positive and non-positive parts, and that the negative part be sufficiently small so as not to compromise the coercivity of the overall operator. Some other recent works have explored extraction of nonlocal models from data, using either PDE-constrained optimization [25] or nonlocal generalizations of physics-informed neural networks [26, 27] to obtain strictly positive kernels, which are thus well-posed. A unique feature of our approach is the extraction of an inequality constraint from Mengesha and Du's theory [24] which allows learning of sign-changing kernels. Approximation of the kernel by Bernstein polynomials allows for a simple implementation of the constraint, guaranteeing extraction of well-posed sign-changing kernels. We will see that in various settings, sign-changing kernels sometimes appear naturally and sometimes do not, and therefore this data-driven derivation may suggest whether such kernels are appropriate for a given application.

Outline of the paper. The paper is organized as follows. To begin, in Section 2 we define an abstract problem characterizing the fitting of nonlocal problems to high-fidelity observations. In Section 3 we review the well-posedness theory for sign-changing kernels that guides the discovery of well-posed nonlocal models and formulate the learning problem as an inequality-constrained optimization problem. Then, in Section 4 we provide a two-stage algorithm to solve the optimization problem efficiently. In Section 5 we illustrate properties of the proposed technique by using manufactured kernels; these results highlight consistency and robustness of our method. In Sections 6, 7, and 8 we consider specific applications such as Darcy's flow, homogenization of a higher-order PDE and a fractional diffusion model. For each application, we highlight the impact of the proposed regression technique. Concluding remarks are reported in Section 9.

2. Abstract problem

We consider in this section a framework for extraction of nonlocal operators which match observations of a given system in an optimal sense. We assume that a quantity of interest u

satisfies the problem

$$\begin{cases} \mathcal{L}_{\mathrm{HF}}[u](x) = f(x) & x \in \Omega, \\ \mathcal{B}u(x) = g(x) & x \in \partial\Omega, \text{ or } u \text{ is periodic}, \end{cases} \qquad (2.1)$$

where $\mathcal{L}_{\mathrm{HF}}$ is a high-fidelity (HF) operator that faithfully represents the system; this can either be a partial differential equation or a fractional-differential operator. The function f denotes a forcing term and operator $\mathcal{B}$ specifies a boundary or exterior condition. For example, when $\mathcal{L}_{\mathrm{HF}}$ is a local operator, $\mathcal{B} = \mathrm{Id}$ corresponds to a Dirichlet boundary condition and $\mathcal{B} = \partial/\partial n$ a Neumann boundary condition. We hypothesize that solutions to this HF problem may be approximated by solutions to an equivalent nonlocal volume-constrained problem of the form

$$\begin{cases} \mathcal{L}_{K}[u](x) = f(x) & x \in \Omega, \\ \mathcal{B}_{I}u(x) = g(x) & x \in \Omega_{I}, \text{ or } u \text{ is periodic}. \end{cases} \qquad (2.2)$$

Here, $\mathcal{L}_K$ is a nonlocal operator parametrized by K, defined below, Ω_I an appropriate nonlocal interaction domain, and B_I the corresponding nonlocal interaction operator specifying a volume constraint. In (2.2), u and f may coincide with the quantities of interest u and f in (2.1), or they may be appropriate "coarse-grained" representations of the same, we postpone a discussion of this multiscale scenario until Section 6.

We seek $\mathcal{L}_K$ as a linear nonlocal operator of the form

$$\mathcal{L}_{K}[u](x) = -\int_{\Omega \cup \Omega_{I}} K(x,y)\left(u(y) - u(x)\right) dy, \qquad (2.3)$$

where K denotes a nonlocal kernel compactly supported on the ball of radius δ centered at x, i.e. $B_{\delta}(x)$. Desired properties of the kernel and details on the analysis of problems associated with operator (2.3) are reported in Section 3. The interaction domain is defined as

$$\Omega_{I} = \{y \in \mathbb{R}^{d} \text{ such that } |y - x| \leq \delta \text{ for some } x \in \Omega\}; \qquad (2.4)$$

a condition for u on Ω_I is required to specify $\mathcal{L}_K u(x)$ for $x \in \Omega$, by (2.3). Thus, an appropriate exterior volume condition on Ω_I replaces the local boundary condition in (2.2). For simplicity we will assume $\mathcal{B}_I$ to be given: in the case of manufactured solutions one may apply the solution as a Dirichlet condition, while in the periodic case it need not be specified.

To learn the operator $\mathcal{L}_K$, we assume that we are given a collection of pairs of forcing

4

terms f_i and corresponding u_i which arise from solutions to (2.1),

$$\mathcal{D} = \{(u_i, f_i)\}_{i=1}^N.$$

These observations will be used to train (2.3). In the simplest setting, $\mathcal{D}$ consists of pairs of forcing terms and corresponding solutions of (2.1). In specific coarse-graining applications, we can apply appropriate post-processing – averaging, for example – to coarsen such "fine-scale" solutions to construct $\mathcal{D}$. We then extract the nonlocal model by finding a kernel such that the action of $\mathcal{L}_K$ most closely maps u_i to f_i.

That is, we solve an optimization problem of the form

$$K^* = \underset{K}{\operatorname{argmin}} \sum_{i=1}^N \left\| \mathcal{L}_K[u_i] - f_i \right\|_{\mathcal{X}}, \tag{2.5}$$

where $\left\| \cdot \right\|_{\mathcal{X}}$ denotes an appropriate norm over Ω.

In doing so, since f_i represents $\mathcal{L}_{\mathrm{HF}}[u_i]$ either exactly or in a post-processed sense, $\mathcal{L}_{K^*}$ best matches the action of $\mathcal{L}_{\mathrm{HF}}$ on the training data $\mathcal{D}$, in the sense of problem (2.5).

In this manner, solution of the problem associated with $\mathcal{L}_{K^*}$ provides a surrogate for the given high-fidelity problem (2.1). We note, however, that there is no reason to expect the associated problem (2.2) to be solvable. Previous works have imposed solvability via imposition of kernel positivity (either by positivity constraint [25, 28] or restricting parameterization to only admit positive kernels [26]), they do not generalize naturally to kernels with negative part. In the next section, we consider a particular class of kernels allowing the addition of constraints to the optimization problem (2.5) to guarantee extraction of well-posed models.

We use this abstract setting as a framework for extracting nonlocal surrogates in a variety of scenarios where nonlocal models are expected to provide advantages. Before specializing to these application-specific settings, we first gather relevant theory regarding the stability and well-posedness of sign-changing nonlocal operators.

3. Data-driven discovery of well-posed nonlocal models

The ultimate goal of the operator regression discussed above is obtain a nonlocal model (2.2) that can be solved for a class of forcing terms f of interest – for example, forcing functions at a suitable coarse scale and with certain regularity properties. However, without a careful definition and parametrization of a feasible set of such nonlocal operators $\mathcal{L}_K$ of the form (2.3), there is no guarantee that regression will recover an operator $\mathcal{L}_K$ that enjoys this property, meaning that when (2.2) is discretized, the resulting linear system will be singular. Such operators cannot be used to extrapolate and generate predictions from new

forcing functions f. Thus, we must avoid learning kernels K^* that solve (2.5), but lead to non-invertible operators $\mathcal{L}_{K^*}$.

The article [24] gives an overview and establishes a generalized framework of well-posed nonlocal problems based on properties of the nonlocal kernel. In this context, well-posedness refers to existence and uniqueness of a solution to a nonlocal problem in an appropriate weak or variational sense, implying that suitable finite-dimensional discretizations of the problem can be solved numerically in a consistent and stable way.

Below, we review the theory of [24] which we will use throughout the article to ensure the discovery of well-posed nonlocal models. This reference discusses two sets of conditions for well-posedness. First, it reviews the case of nonlocal operators with *nonnegative* kernels, reviewed in Section 3.1. Next, it weakens this condition and establishes well-posedness for a class of sign-changing kernels satisfying conditions on the negative part of the kernel function, reviewed in Section 3.2. We shall consider operator regression in the latter class of kernels. Our purpose for this is two-fold: first, this more general class of kernels is, in principle, more expressive and able to provide better fit to data. Second, we aim to study the utility of this sign-changing kernels for nonlocal modeling. In other words, we study the question: is there an advantage to using sign-changing kernels, as opposed to nonnegative kernels, in fitting nonlocal models to data?

The conditions for well-posedness lead to constraints for the operator regression optimization problem. We define our parametrization of the unknown kernel in Bernstein polynomials and state the constrained optimization problem that arises in Section 3.3.

3.1. Well-posedness with nonnegative kernels

In this section we introduce a standard nonlocal model with a nonnegative kernel and recall conditions for the well-posedness of associated diffusion problems. Let $\Omega \subset \mathbb{R}^d$ be a connected bounded domain with sufficiently smooth boundary and let Ω_I be its interaction domain defined as in (2.4). We denote their union by $\widehat{\Omega} = \Omega \cup \Omega_I$. We define the action of a nonlocal diffusion operator $\mathcal{L}_\rho$ on a function u as

$$\mathcal{L}_\rho u(x) = -2 \int_{\widehat{\Omega}} \rho(|y - x|)\, (u(y) - u(x))\, dy, \tag{3.1}$$

where the nonnegative and compactly supported kernel $\rho = \rho(|\xi|)$ is such that

$$|\xi|^2 \rho(|\xi|) \in L^1_{\text{loc}}(\mathbb{R}^d), \text{ and}$$
$$\exists\, \sigma > 0 \text{ such that } (0, \sigma) \subset \text{supp}(\rho). \tag{3.2}$$

To study existence and uniqueness of solutions to the equation

$$\mathcal{L}_\rho u = b, \tag{3.3}$$

we define the space $S(\widehat{\Omega})$ associated with $\mathcal{L}_\rho$ as the typical energy space considered for nonlocal problems [15, 24, 29, 30], i.e.

$$S(\widehat{\Omega}) = \left\{ u \in L^2(\widehat{\Omega}) : \int_{\widehat{\Omega}} \int_{\widehat{\Omega}} \rho(|y-x|)|u(y) - u(x)|^2 dy dx < \infty \right\}.$$

Note that $S(\widehat{\Omega})$ arises naturally from the symmetric bilinear form

$$
\begin{aligned}
(\mathcal{L}_\rho u, w)_{L^2(\widehat{\Omega})} &= -2 \int_{\widehat{\Omega}} w(x) \int_{\widehat{\Omega}} \rho(|y-x|)\,(u(y) - u(x))\,dy dx \\
&= \int_{\widehat{\Omega}} \int_{\widehat{\Omega}} \rho(|y-x|)\,(u(y) - u(x))\,(w(y) - w(x))\,dy dx.
\end{aligned}
$$

We denote by $\mathcal{S}'$ the dual space of $\mathcal{S}$. Let V be a closed subspace of $L^2(\widehat{\Omega})$ that contains the constant function $u \equiv 0$ and no other constant functions, we define the solution space as $\mathcal{V}_s = V \cap \mathcal{S}(\widehat{\Omega})$. According to Corollary 1 of [24], if ρ satisfies (3.2), there exists a coercivity constant $\kappa = \kappa(\rho, V, \widehat{\Omega})$ of $\mathcal{L}_\rho : \mathcal{S} \to \mathcal{S}'$ such that

$$\|u\|^2_{L^2(\widehat{\Omega})} \leq \kappa\,(\mathcal{L}_\rho u, u)_{L^2(\widehat{\Omega})}, \qquad \text{for all } u \in V \cap S(\widehat{\Omega}), \tag{3.4}$$

where $(\cdot, \cdot)_{L^2(\cdot)}$ denotes the L^2 inner product. By Lemma 3.1 of [24], this result guarantees that there exists a unique $u \in \mathcal{V}_s$ such that, for $b \in \mathcal{S}'$ and for all $w \in \mathcal{V}_s$

$$(\mathcal{L}_\rho u, w)_{L^2(\Omega)} = (b, w)_{L^2(\Omega)}, \tag{3.5}$$

Equation (3.5) is the weak form of (3.3).

3.2. Well-posedness for sign-changing kernels

In this section we consider the type of sign-changing kernels studied by [24]. These kernels define operators of the form

$$\mathcal{L}_K = \mathcal{L}_\rho + \lambda \mathcal{L}_g, \tag{3.6}$$

where $\mathcal{L}_\rho$ is the operator defined in (3.1) and the nonlocal operator $\mathcal{L}_g$ on the right-hand side is defined as

$$\mathcal{L}_g u(x) = -2 \int_{\widehat{\Omega}} g(|y-x|)\,(u(y) - u(x))\,dy.$$

Thus, we consider integral operators of the form

$$\mathcal{L}_K u(x) = -\int_{\widehat{\Omega}} K(x, y)\left(u(y) - u(x)\right) dy,$$

with kernel

$$K(x, y) = 2\rho(|y - x|) + 2\lambda g(|y - x|). \tag{3.7}$$

The function ρ is assumed to satisfy the conditions (3.2) for well-posedness of the problem (3.3), so that the associated bilinear form is coercive as in (3.4) with coercivity constant κ. The idea of [24] is to think of $\mathcal{L}_K$ as a small perturbation of $\mathcal{L}_\rho$ by $\lambda\mathcal{L}_g$. Then, for well-posedness of equations of the form

$$\mathcal{L}_K u = b, \tag{3.8}$$

we need the contribution of the perturbation $\mathcal{L}_g$ not to compromise the coercivity of the variational form associated with $\mathcal{L}_\rho$, so that $\mathcal{L}_K$ will also induce a coercive, hence well-posed, variational problem.

We now summarize the conditions on g and λ that guarantee well-posedness of the perturbed system (3.6). From Lemma 4.2, Theorem 4.3, and Corollary 2 of [24], have that if g is a compactly supported locally integrable radial function, and if g and λ satisfy

$$|\lambda| < \frac{1}{2\kappa\left(\|g\|_{L^1(\mathbb{R}^d)} + \|G\|_{L^\infty(\widehat{\Omega})}\right)}, \tag{3.9}$$

with $G(x) = \int_{\widehat{\Omega}} g(|y - x|) dy$ and κ defined as in (3.4), then, for $b \in \mathcal{V}_s'$, the weak form of equation (3.8)

$$(\mathcal{L}_K u, w)_{L^2(\Omega)} = (b, w)_{L^2(\Omega)}, \quad \forall w \in \mathcal{V}_s,$$

has a unique solution $u \in \mathcal{V}_s$. This is the analogue of (3.5) for the equation (3.8).

In the following section, we will parametrize ρ and λ in coefficients of a polynomial expansion, making λ in (3.7) a redundant parameter that scales the coefficients. Writing $h = \lambda g$, multiplying (3.9) through by the factor $\left(\|g\|_{L^1(\mathbb{R}^d)} + \|G\|_{L^\infty(\widehat{\Omega})}\right)$ and pulling λ inside the norms, the setup of (3.7) with condition (3.9) can be written as

$$\begin{cases} K(x, y) = 2\rho(|y - x|) + 2h(|y - x|), \\ H(x) = \int_{\widehat{\Omega}} h(|y - x|) dy, \\ \|h\|_{L^1(\mathbb{R}^d)} + \|H\|_{L^\infty(\widehat{\Omega})} < \frac{1}{2\kappa}. \end{cases} \tag{3.10}$$

8

Accordingly, we also define $\mathcal{L}_h = \lambda \mathcal{L}_g$, so that the operator $\mathcal{L}_K$ in (3.6) can be written as $\mathcal{L}_K = \mathcal{L}_\rho + \mathcal{L}_h$.

3.3. Sign-changing kernel regression

With the purpose of applying the abstract formulation introduced in Section 2 to the discovery of kernels described in Section 3, we introduce a representation of the kernel as a linear combination of Bernstein basis polynomials. Notably, we assume that K is radial, i.e., that $K(x, y)$ is a function of $|\frac{x-y}{\delta}|$, and expand in polynomials in this single variable. Specifically, we parametrize (3.7) by defining

$$K(x, y) = \sum_{m=0}^{M} \frac{C_m}{\delta^{d+2}} B_{m,M}\left(\left|\frac{x - y}{\delta}\right|\right) + \sum_{m=0}^{M} \frac{D_m}{\delta^{d+2}} B_{m,M}\left(\left|\frac{x - y}{\delta}\right|\right), \tag{3.11}$$

where the Bernstein basis functions are defined as

$$B_{m,M}(x) = \binom{M}{m} x^m (1 - x)^{M-m}.$$

This corresponds to setting

$$2\rho(|\xi|) = \sum_{m=0}^{M} \frac{C_m}{\delta^{d+2}} B_{m,M}\left(\left|\frac{\xi}{\delta}\right|\right), \quad \text{and} \quad 2h(|\xi|) = \sum_{m=0}^{M} \frac{D_m}{\delta^{d+2}} B_{m,M}\left(\left|\frac{\xi}{\delta}\right|\right) \tag{3.12}$$

in (3.10), where the coefficients of the linear combinations are unknown. The scaling of C_m's and D_m's is standard and has been used in the literature (see, e.g. [31]) to guarantee convergence of nonlocal diffusion operators to Δ, the classical Laplacian, as $\delta \to 0$.

For $\mathbf{C} \in \mathbb{R}^{M+1}$, $\mathbf{D} \in \mathbb{R}^{M+1}$, the kernel discovery problem can be stated as

$$\begin{cases} \min_{\mathbb{R}^{2M+2}} L\left(\begin{bmatrix} \mathbf{C} \\ \mathbf{D} \end{bmatrix}\right) \\ \text{subject to:} \quad \mathbf{C} \geq \mathbf{0}, \quad N(\mathbf{D}) \leq [2\kappa(\mathbf{C})]^{-1}, \end{cases} \tag{3.13}$$

9

where

$$
\begin{cases}
L\left(\begin{bmatrix}\mathbf{C}\\ \mathbf{D}\end{bmatrix}\right) = \dfrac{1}{N}\sum_{i=1}^{N}\|(\mathcal{L}^{\mathbf{C}}+\mathcal{L}^{\mathbf{D}})[u_i]-f_i\|_{\mathcal{X}}^2\\[3mm]
N(\mathbf{D}) = \dfrac{1}{2}\left\|\displaystyle\sum_{m=0}^{M}\dfrac{D_m}{\delta^{d+2}}B_{m,M}\left(\left|\dfrac{x}{\delta}\right|\right)\right\|_{L^1(\mathbb{R}^d)} + \dfrac{1}{2}\left\|\displaystyle\sum_{m=0}^{M}\dfrac{D_m}{\delta^{d+2}}\int_{\mathbb{R}}B_{m,M}\left(\left|\dfrac{y-x}{\delta}\right|\right)dy\right\|_{L^\infty(\widehat{\Omega})}\\[3mm]
\kappa(\mathbf{C}) \text{ is a constant satisfying (3.4).}
\end{cases}
$$

$$(3.14)$$

Above, $\mathcal{L}^{\mathbf{C}}$ denotes $\mathcal{L}_\rho$, $\mathcal{L}^{\mathbf{D}}$ denotes $\mathcal{L}_h$, following the parametrization (3.12), and we delay specification of the norm $\|\cdot\|_{\mathcal{X}}$ until the following section. Numerical optimization of this problem is discussed in Section 4.

4. Algorithm

The problem (3.13) is expected to be nonconvex and susceptible to many local minima. At the same time, while the constraint $\mathbf{C}\geq\mathbf{0}$ is easier to enforce, the second constraint involving $N(\mathbf{D})$ and $\kappa(\mathbf{C})$ in (3.14) is highly nonlinear in both the left-hand and right-hand sides, posing a significant challenge. For numerical efficiency, we have followed a "two-stage" strategy for numerical optimization.

We first find a set of $\mathbf{C}^*$ representing an initial fit to the data using only the nonnegative kernel ρ as in (3.12), ignoring the term h and avoiding the nonlinear constraint. We minimize the loss

$$
L_1(\mathbf{C}) = \frac{1}{N}\sum_{i=1}^{N}\left\|\sum_{m=0}^{M}\frac{\text{ReLU}(C_m)}{\delta^3}\left[\int_{\widehat{\Omega}}B_{m,M}\left(\left|\frac{y-x}{\delta}\right|\right)(u_i(x)-u_i(y))dy\right]_{\Delta}-f_i(x)\right\|_{\mathcal{X}}^2 \quad (4.1)
$$

Here we use the notation $\left[\int\cdot\,dy\right]_{\Delta}$ to denote the discrete approximation of a nonlocal kernel by applying quadrature. For the purposes of this work, we will use Silling's one-point quadrature [32], popular in particle discretizations of peridynamics. We note however that any choice of quadrature for discretization of the integral in the strong form [33, 34], or basis for the discretization of the weak form, may be applied here; see [35] and references therein. For a given nonlocal operator $\mathcal{L}_K$ and basis functions $\mathbf{\Phi}=[\phi_p]_{p=1}^{P}$ and corresponding degrees of freedom $\mathbf{w}=[w_p]_{p=1}^{P}$ the discretization provides mass and nonlocal stiffness matrices

$$
M_{ij}=\left[\int_{\Omega}\phi_i(x)\phi_j(x)\,dx\right]_{\Delta} \quad \text{and} \quad S_{ij}=\left[\int_{\Omega}\mathcal{L}_K[\phi_i(x)]\phi_j(x)dx\right]_{\Delta}. \quad (4.2)
$$

Note the presence of the ReLU function in L_1. We minimize L_1 using the Adam optimizer

[36], also mapping

$$\mathbf{C} \mapsto \text{ReLU}(\mathbf{C})$$

after each step of gradient descent. This leaves L_1 invariant but ensures that the sequence and local minimum produced by gradient descent satisfies $\mathbf{C} \geq 0$. Thus, the simpler constraint $\mathbf{C} \geq 0$ is hard-coded into the loss function and algorithm, allowing us to obtain $\mathbf{C}^*$ by applying the Adam algorithm to the unconstrained optimization problem.

We then compute the coercivity constant $\kappa(\mathbf{C}^*)$ corresponding to the operator $\mathcal{L}^{\mathbf{C}^*}$. For $\mathbf{w}, S$, and M consistent with the choice of discretization $\left[\int \cdot\, dy\right]_\Delta$, we solve the following generalized eigenvalue problem

$$[\kappa(\mathbf{C}^*)]^{-1} = \min_{\mathbf{w}} \frac{\mathbf{w}^T S \mathbf{w}}{\mathbf{w}^T M \mathbf{w}}. \tag{4.3}$$

This computation happens only *once* in our algorithm, so that $\mathbf{C}^*$ and $\kappa(\mathbf{C}^*)$ are fixed in the next stage.

In the second stage, we correct the initial fit given by $\mathcal{L}^{\mathbf{C}^*}$ by finding $\mathcal{L}^{\mathbf{D}^*}$ subject to the second constraint in (3.13), in which the right-hand side $[2\kappa(\mathbf{C}^*)]^{-1}$ of the constraint is fixed. We apply the augmented Lagrangian method [37, 38] to solve for $\mathbf{D}$ under the constraint condition, using the function

$$H(\mathbf{D}, \theta) = \frac{1}{2\kappa(\mathbf{C}^*)} - N(\mathbf{D}) - \theta^2$$

$$= \frac{1}{2\kappa(\mathbf{C}^*)} - \frac{1}{2} \sum_{m=0}^{M} \frac{D_m}{\delta^{d+2}} \left[\int_{\mathbb{R}^d} B_{m,M}\left(\left|\frac{x}{\delta}\right|\right) dy\right]_\Delta$$

$$- \frac{1}{2} \sup_{\widehat{\Omega}} \left|\sum_{m=0}^{M} \frac{D_m}{\delta^{d+2}} \left[\int_{\mathbb{R}^d} B_{m,M}\left(\left|\frac{y-x}{\delta}\right|\right) dy\right]_\Delta\right| - \theta^2.$$

Here, θ is the slack variable arising from the inequality constraint. We then apply the Adam optimizer to the penalized loss function

$$L_2(\mathbf{D}, \theta) = \frac{1}{N} \sum_{i=1}^{N} \|\mathcal{L}^{\mathbf{C}^*+\mathbf{D}}[u_i] - f_i\|_{\mathcal{X}}^2 + \lambda H(\mathbf{D}, \theta) + \frac{\mu}{2} H^2(\mathbf{D}, \theta) \tag{4.4}$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left\|\sum_{m=0}^{M} \frac{C_k + D_k}{\delta^3} \left[\int_{\widehat{\Omega}} B_{m,M}\left(\left|\frac{y-x}{\delta}\right|\right)(u_i(x) - u_i(y)) dy\right]_\Delta - f_i(x)\right\|_{\mathcal{X}}^2$$

$$+ \lambda H(\mathbf{D}, \theta) + \frac{\mu}{2} H^2(\mathbf{D}, \theta),$$

adjusting the parameters λ and μ as described in Algorithm 1. We denote the minimum by

$\mathbf{D}^*$.

Although this two-stage optimization algorithm is numerically efficient, there is a degree of overconstraining arising from a single computation of the coercivity constant κ. That is, although $\mathbf{C}^*$ is a local minimum of L_1 and $\mathbf{D}^*$ is a local minimum of L_2, $(\mathbf{C}^*, \mathbf{D}^*)$ is not necessarily a local minimum of L.

A more advanced algorithm could involve iterated computation of the coercivity constraint κ. In our examples below, however, we found that the two-stage algorithm is sufficient and that there was no significant benefit to more complex and expensive optimization algorithms. Therefore, we have used the two-stage algorithm throughout.

Algorithm 1 Nonlocal kernel regression

1: Initialize $C_m^{(0)} \sim \mathcal{U}(0,1)$.
2: Obtain $\mathbf{C}^*$ as a local minimum of $L_1(\mathbf{C})$, using the Adam optimizer while updating $\mathbf{C} \leftarrow \mathrm{ReLU}(\mathbf{C})$ after each step of gradient descent.
3: Select basis $\{w_p(x)\}$, $p = 1, ..., P$, and assemble the matrices M and S in (4.2).
4: Solve the generalized eigenvalue problem (4.3) for $\kappa(\mathbf{C}^*)$.
5: Initialize $D_m^{(0)} \sim \mathcal{U}\left(-\frac{1}{\sqrt{M+1}}, \frac{1}{\sqrt{M+1}}\right)$ and $\theta^{(0)} = 1$.
6: Set STEP_MAX $= 100$, $\lambda = 0$, $\mu = 1$, $s = 1$, $\rho = 10$, $c = 1/4$, $\epsilon = 10^{-8}$.
7: **while** $s \leq$ STEP_MAX: **do** ▷ Perform Augmented Lagrangian Algorithm for $\mathbf{D}$
8:    Solve the unconstrained optimization problem

$$(\mathbf{D}^{(s)}, \theta^{(s)}) = \underset{\mathbf{D}, \theta}{\mathrm{argmin}}\, L_2(\mathbf{D}, \theta).$$

9:    **if** $H(\mathbf{D}^{(s)}, \theta^{(s)}) \leq \epsilon$ **then**
10:        Stop.
11:    **else**
12:        **if** $H(\mathbf{D}^{(s)}, \theta^{(s)}) \geq cH(\mathbf{D}^{(s-1)}, \theta^{(s-1)})$ **then**
13:            Update penalty $\mu \leftarrow \rho\mu$.
14:            **if** $\mu \geq 10^{20}$ **then**
15:                Stop.
16:        **else**
17:            Update Lagrange multiplier $\lambda \leftarrow \lambda + \mu H(\mathbf{D}^{(s)}, \theta^{(s)})$.
18:    Update the iteration number $s \leftarrow s + 1$.
19: $\mathbf{D}^* = \mathbf{D}^{(s)}$.

In applying Algorithm 1, we run the Adam optimizer (in PyTorch) using a batch size of 100 and learning rate 5e-3 throughout the article. We run until the loss stagnates, indicating that a stationary point has been reached. This was typically between 200 and 500 epochs for the first stage of the algorithm (to find $\mathbf{C}^*$) and 10 epochs for the second stage (to find $\mathbf{D}^*$) per iteration of the augmented Lagrangian method. In general we will take at least $O(1,000)$

samples for the purposes of training, to ensure we are well into the regime of having sufficient data. Throughout, we use the norm

$$\|f\|_{\mathcal{X}} = \sqrt{\frac{1}{\#\mathcal{X}} \sum_{x_k \in \mathcal{X}} f(x_k)^2}$$

in (4.1) and (4.4), for a discrete collection $\mathcal{X}$ of points in Ω. We specify the collection $\mathcal{X}$ in the sections below.

5. Tests with Manufactured Kernels

In this section, we give the first illustration of the method described above by training on data $\mathcal{D}_{\text{train}} = \{(u_i, f_i)\}_{i=1}^{N}$ generated from a nonlocal equation of the form (2.2) with periodic boundary conditions, for a manufactured kernel K_{man} that is given *a priori*. This allows us to validate the approach and quantify its ability to handle nonnegative and sign-changing kernels, before moving to noisy data in subsequent sections.

We generate 50,000 training pairs (u_i, f_i) in the following way. First we randomly generate the coefficients of a Fourier series for u_i as

$$\widehat{u}_k \sim \exp(-\alpha k^2)\xi, \quad \xi \sim \mathcal{U}[0, 1], \tag{5.1}$$

where $\mathcal{U}[0, 1]$ denotes the uniform distribution on $[0, 1]$ and $\alpha = 0.1$. Then, u_i is given by

$$u_i = \sum_{k=0}^{100} \widehat{u}_k \cos(2\pi k x / L) \tag{5.2}$$

and f_i is obtained from (2.1) via a numerical Fourier transform. We train the kernel using $\mathcal{X}$ to be 101 equidistant points in $[0, 1]$.

The learned kernel K^* is validated on a test set consisting of a different set of samples $\mathcal{D}_{\text{test}} = \{(u_j, f_j)\}_{j=1}^{N}$ generated as above, with $N = 10,000$. We test two types of basis expansions in (3.11) for learning K: a linear kernel ($M = 1$) and a quadratic kernel ($M = 2$).

For the case of a linear manufactured kernel

$$K_{\text{man}}(x, y) = \frac{4}{\delta^3} \left| \frac{y - x}{\delta} \right| \mathcal{X}_{[0,1]} \left(\left| \frac{y - x}{\delta} \right| \right), \tag{5.3}$$

used to generate data, prediction using $\mathcal{L}_{K^*}$ is shown for two test pairs (u, f) in Figure 1 for regression using a linear kernel, and Figure 2 for regression using a quadratic basis. In these first examples, we only learned the nonnegative coefficients $\mathbf{C}$, effectively setting $\mathbf{D} = 0$

13

and not performing the second stage of the algorithm. This verifies that given data from a nonnegative manufactured kernel, the algorithm can provide a good fit with a nonnegative learned kernel.
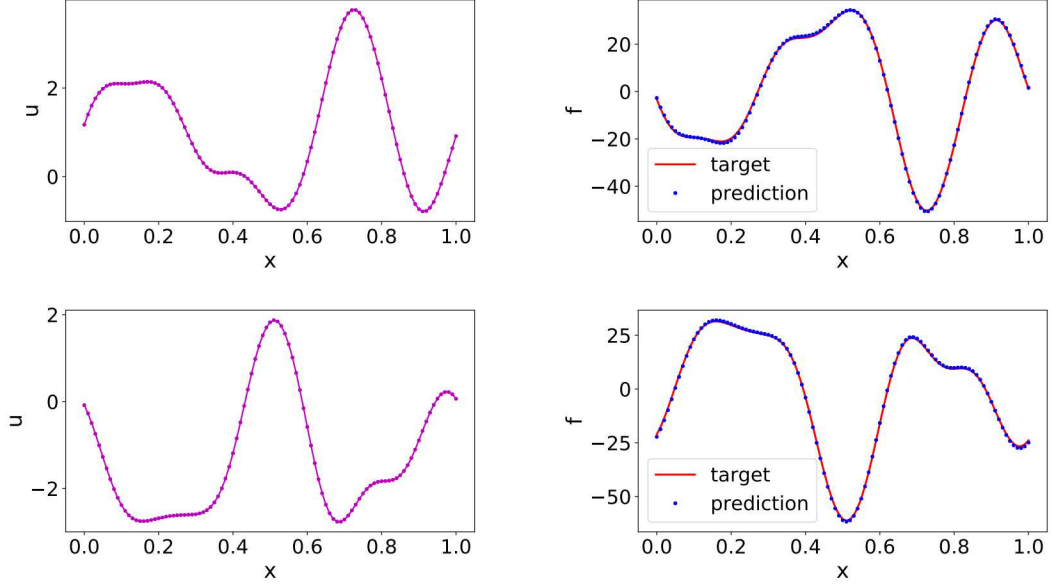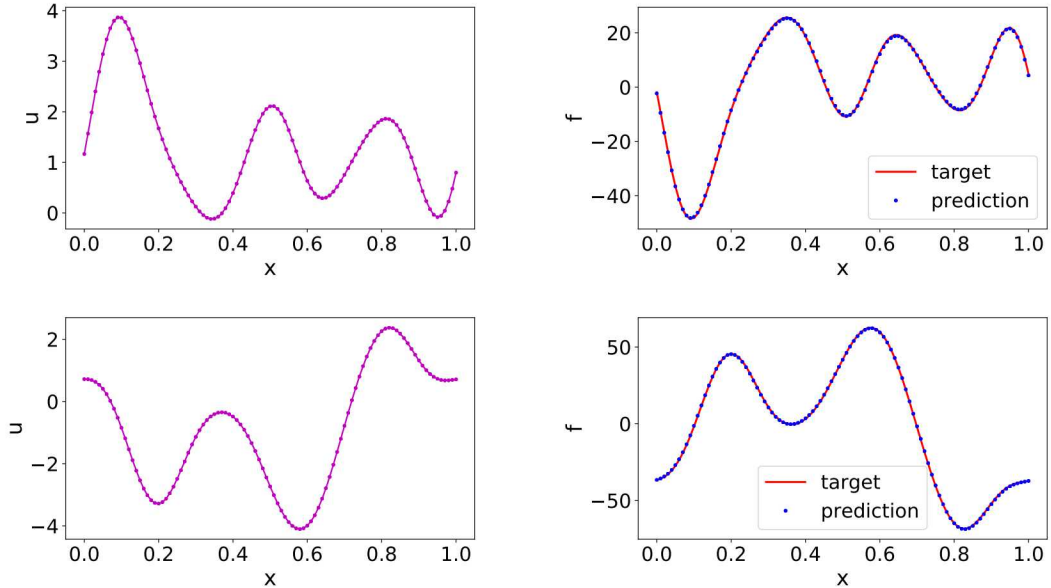

Figure 1: Comparison of the prediction $\mathcal{L}_{K^*}[u]$ and target f when generating data from the manufactured kernel (5.3) and fitting a nonnegative linear basis polynomial kernel K^*. The left column illustrates u, whereas the right column f (target) and $\mathcal{L}_{K^*}[u]$ (prediction), for a sample of $(u, f) \in \mathcal{D}_{\text{test}}$ in each row.

Figure 2: Comparison of the prediction $\mathcal{L}_{K^*}[u]$ and target f when generating data from the manufactured kernel (5.3) and fitting a nonnegative quadratic basis polynomial kernel K^*. The left column illustrates u, whereas the right column f (target) and $\mathcal{L}_{K^*}[u]$ (prediction), for a sample of $(u, f) \in \mathcal{D}_{\text{test}}$ in each row.

In the next example, we fit data generated from the sign-changing kernel

$$K_{\mathrm{man}}(x, y) = \frac{21.4615}{\delta^3} \cos\left(\frac{3\pi|y - x|}{5\delta}\right). \tag{5.4}$$

Here, we apply the second stage of Algorithm 1, learning both $\mathbf{C}$ and $\mathbf{D}$. We use this example to explore whether learning a kernel K^* by fitting the action of $\mathcal{L}_K$ on the data $\mathcal{D}_{\mathrm{train}}$ can be expected to reproduce the kernel K_{man}. Of course, for a general kernel K_{man}, this can only be expected for increasing basis order M of the polynomial expansion in (3.11). We generate two sets of training and test data. The first set, which we refer to as "low-frequency", consists of 50,000 samples of the form (5.2) with distribution (5.1), just as for the previous example. The second set, which we refer to as "high-frequency", consists of $25,000$ of such low-frequency samples and $25,000$ samples (u_i, f_i) where

$$u_i = \xi_1 \sin(2\pi k_1 x/L) + \xi_2 \cos(2\pi k_2 x/L), \tag{5.5}$$

for $\xi_1, \xi_2 \sim \mathcal{U}[0, 1]$ and k_1 and k_2 being random integers sampled uniformly in $[5, 15]$, and with f_i computed using a numerical Fourier transform.

We learn kernels K of the form (3.11) with degrees M of the basis increasing 2 to 20, for both data sets.

The plots of training and validation losses for both sets of data are shown in Figure 3. This figure shows no benefit to increasing the basis order past 11 for the case of low-frequency data, in contrast to the high-frequency test data for which both losses improve through order 20. This test illustrates that the choice of basis order M should take into account the frequency of the training and test data, an issue which will arise in Section 6 below. Next, to study the question of reproducing the manufactured kernel K_{man} with the learned kernel K^*, we compare K_{man} to K^* first for the low-frequency data and varying basis order M in Figure 4. We show both the nonnegative part 2ρ and the full kernel $2\rho + 2h$ to illustrate the contribution of $2h$.
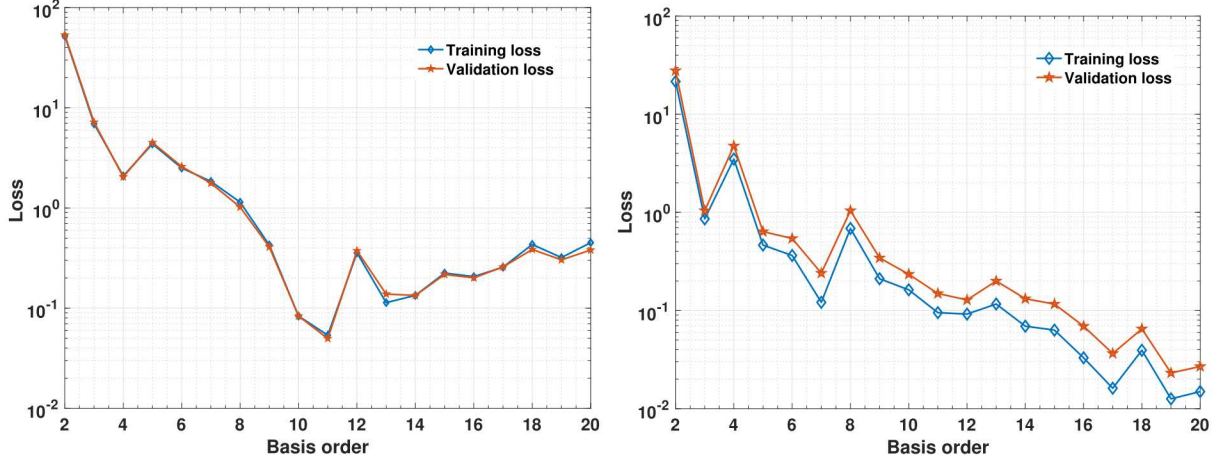
Figure 3: *Left:* The training loss and validation loss versus basis order for low-frequency data generated using the manufactured kernel K_{man} (5.4). *Right:* The training loss and validation loss versus basis order for high-frequency data generated using the manufactured kernel K_{man} (5.4).
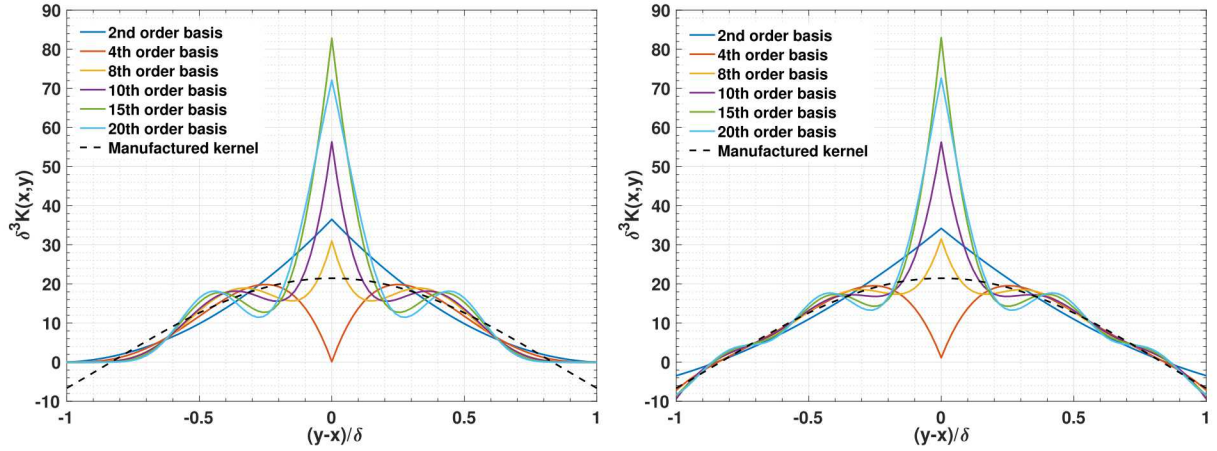

Figure 4: *Left:* Comparison of the nonnegative part $2\rho(x,y)$ of the kernel K^* trained with low-frequency data and the manufactured kernel K_{man} (5.4). *Right:* Comparison of the full kernel $K^* = 2\rho(x,y) + 2h(x,y)$ trained with low-frequency data and the manufactured kernel K_{man} (5.4). Note the improved fit of the negative tails of the kernel.

While the kernel is approximated by $2\rho + 2h$ for larger $|y - x|/\delta$ and sufficiently high-order basis, Figure 4 demonstrates that the manufactured kernel is not recovered for small $|x - y|/\delta$, implying that the action of the operator ρ is insensitive to the kernel shape for small $|x - y|/\delta$ for training data f_i of the form (5.2) with (5.1). We hypothesize this is due to lack of higher frequencies in f_i and the resulting solution u_i. By repeating this experiment in Figure 5 for the high-frequency data (5.5), we obtain evidence for this hypothesis, as a closer fit is obtained for small $|x - y|/\delta$ than for the low-frequency training data in Figure 4 as the basis increases.

These experiments suggest that our algorithm cannot be expected to recover exactly

16

the kernel K_{man} from the data $\mathcal{D}_{\text{train}}$. Moreover, resolution of the kernel in the vicinity of $|x - y|/\delta = 0$ is aided by the incorporation of higher frequency training data. Perhaps unsurprisingly, for such a sign-changing kernel, incorporation of $\mathbf{D}$ improves reproduction of K_{man}.
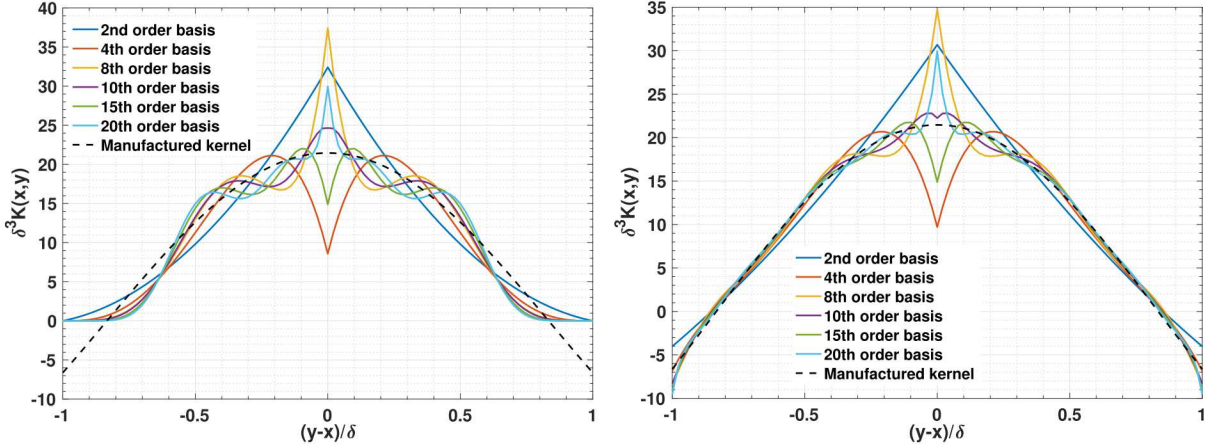

Figure 5: *Left:* Comparison of the nonnegative part $2\rho(x, y)$ of the kernel K^* trained with high-frequency data and the manufactured kernel K_{man} (5.4). *Right:* Comparison of the full kernel $K^* = 2\rho(x, y) + 2h(x, y)$ trained with high-frequency data and the manufactured kernel K_{man} (5.4). Note the improved fit of the negative tails of the kernel.

6. Nonlocal coarse-graining of Darcy flow

We next consider extraction of a nonlocal diffusion model by coarse-graining local solutions of Darcy's equation, a problem arising in subsurface flows through porous media [39, 40]. In traditional inverse modeling and subsurface applications, the inference of an effective permissivity from data is a canonical problem motivating the development of multiscale and homogenization methods [41–44]. Some works have argued that nonlocal models converging to Darcy's equations in the limit $\delta \to 0$ provide an improved description of multiscale transport [18, 45]. We consider a simple 1D problem to demonstrate how nonlocal kernels may be derived from data, illustrating viability of data-driven multiscale nonlocal models for such processes. We numerically homogenize a repeated microstructure of length-scale $2L$ using a nonlocal model with support lengthscale δ. We will see that in this nonlocal context, the choice of coarse-graining lengthscale relative to $2L$ is tied to the effectiveness of the homogenization, similar to the local homogenization setting [46].

We assume a periodic domain $\Omega = [0, 10]$, constituted of alternating subdomains of width L and piecewise constant diffusivity $\kappa_1 = 1$, $\kappa_2 = 4$ (Figure 6). Further, we use a P1 nodal finite element solver to generate a collection of solutions $U = \{u_i\}_{i=1}^N$ to the following high-
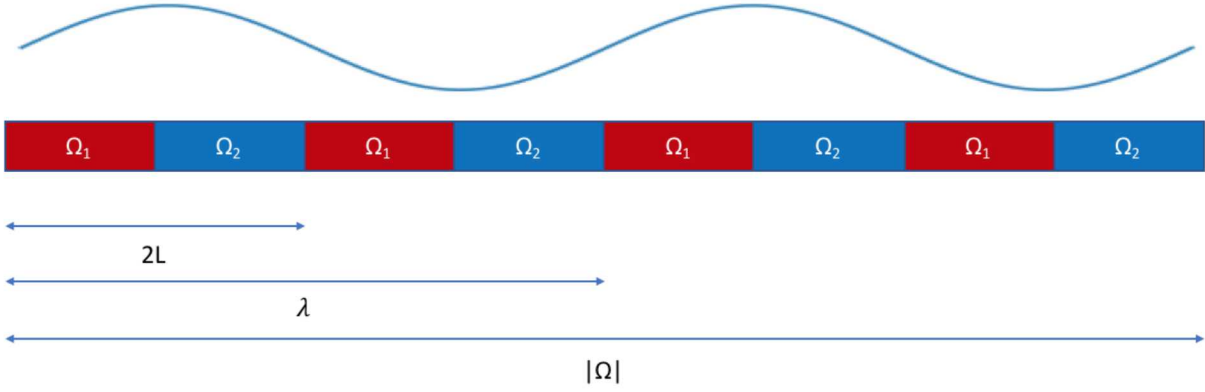
Figure 6: Problem setup and relevant lengthscales for the Darcy coarse-graining problem. Periodic microstructure length is denoted $2L$, while forcing used during testing/training corresponds to $\sin(\lambda\,2\pi x)$, and entire domain has measure $|\Omega|$. We will observe that choice of λ used during test and training must be sufficiently large relative to $2L$ to obtain good fit for data-driven nonlocal model.

fidelity problem

$$\nabla \cdot \mathbf{F}_i = \sin(2\pi x \lambda_i) := f_i$$
$$\mathbf{F}_i = -\kappa(x)\nabla u_i.$$

Here, λ_i denotes a wavelength sampled from the uniform distribution $\lambda_i \sim \mathcal{U}([\lambda_{\min}, |\Omega|])$. We omit details regarding the finite element solution, noting that we use 16 elements per subdomain, which was chosen to ensure sufficient resolution of training data, and take $N = 1000$.

We first coarse-grain U into a collection of averaged solutions $\overline{U} = \{\overline{u}_i\}_{i=1}^N$, where we define $\overline{u}_i$ (and likewise $\overline{f}_i$) as piecewise constant over a given interval $x \in [2Ln, 2L(n+1)]$,

$$\overline{u}_i(x) = \frac{1}{2L} \int_{2Ln}^{2L(n+1)} u_i(y)\, dy,$$

for a given domain n. In this manner, we obtain a representation of the local solution homogenized over the periodic microstructure. Finally, we apply Algorithm 1 to find a nonlocal kernel consistent with $\overline{U}$. That is, we apply the algorithm with

$$\mathcal{D}_{\text{train}} = \{(\overline{u}_i, \overline{f}_i)\}_{i=1}^N.$$

A key feature of this data is the relative magnitude of the wavelength λ_i and the microstructure L; similar to traditional homogenization we require L/λ_i sufficiently small to treat the material as a homogeneous continuum. In Figure 7 we present solutions to a pair
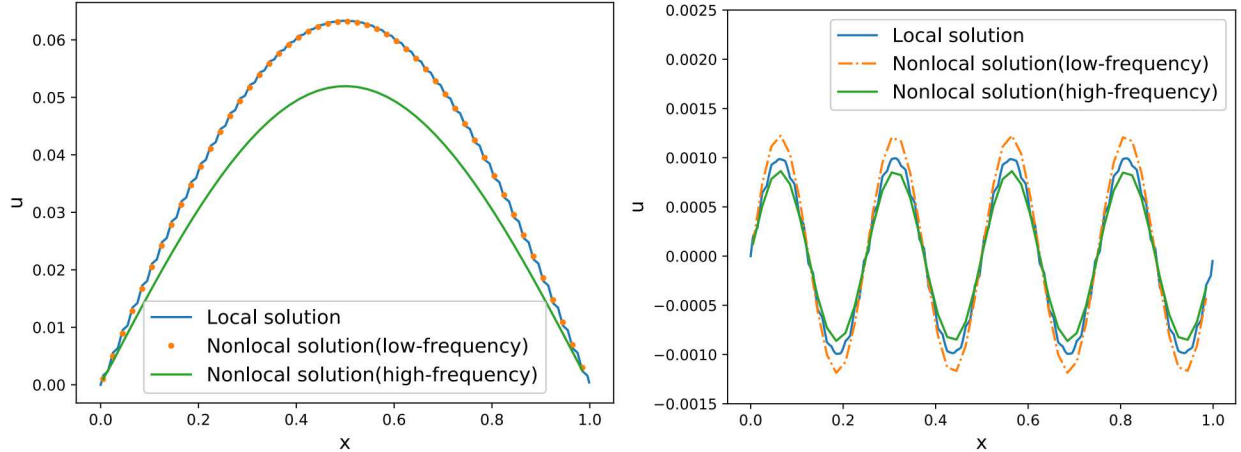
Figure 7: Representative solutions to the extracted kernels, for $\delta = 8L$, testing against low-frequency *(left)* and high-frequency *(right)* forcing. We note that for low-frequency forcing, the solution is primarily smooth, and captured well by models extracted from low-frequency data. For high-frequency forcing, the local solution has artifacts from the microstructure, preventing extraction of an effective model using either high- or low-frequency training data.

of models with $\lambda_{\min} = 1/16$ and 1, both for $\delta = 8L$. When considering a low-frequency solution ($\mathcal{L}[u] = \sin \pi x$), the model trained on low-frequency data matches its local counterpart well, while for a high-frequency problem ($\mathcal{L}[u] = \sin 8\pi x$) neither model performs well. This suggests that $L/\lambda_{\min}$ must be sufficiently small in the training data to prevent the model from overfitting to high-frequency data. When applying the model, it will only agree with the local solution over sufficiently large lengthscales.

Motivated by this observation, we perform a study sweeping over the choice of $\lambda_{\min}$, both during training and testing, for $\delta \in \{4L, 8L, 16L\}$. The results of the study are presented in Figure 8, with the trend that smaller choice of δ provides generally better results. The extracted kernels associated with this study are presented in Figure 9. We observe that for different δ, the resulting kernel may have qualitatively different shape, and that for high-frequency training (which tests poorly) the kernel shifts to a different shape. Most notably, for this problem we observe no discernable negative part in the kernel, despite the fact that the algorithm allows for it. This suggests that, at least for this diffusion process, it is not necessary to introduce negativity into the kernel to obtain good fit; similar nonnegative kernels were obtained in [18] as an upscaling of a pore network.

7. Extraction of nonlocal sign-changing kernels from high-order PDEs

We now seek an application where we expect to extract a sign-changing kernel. Weckner and Silling summarized several works where authors augment second-order elliptic problems with high-order derivatives to accommodate the high-frequency response of a given material
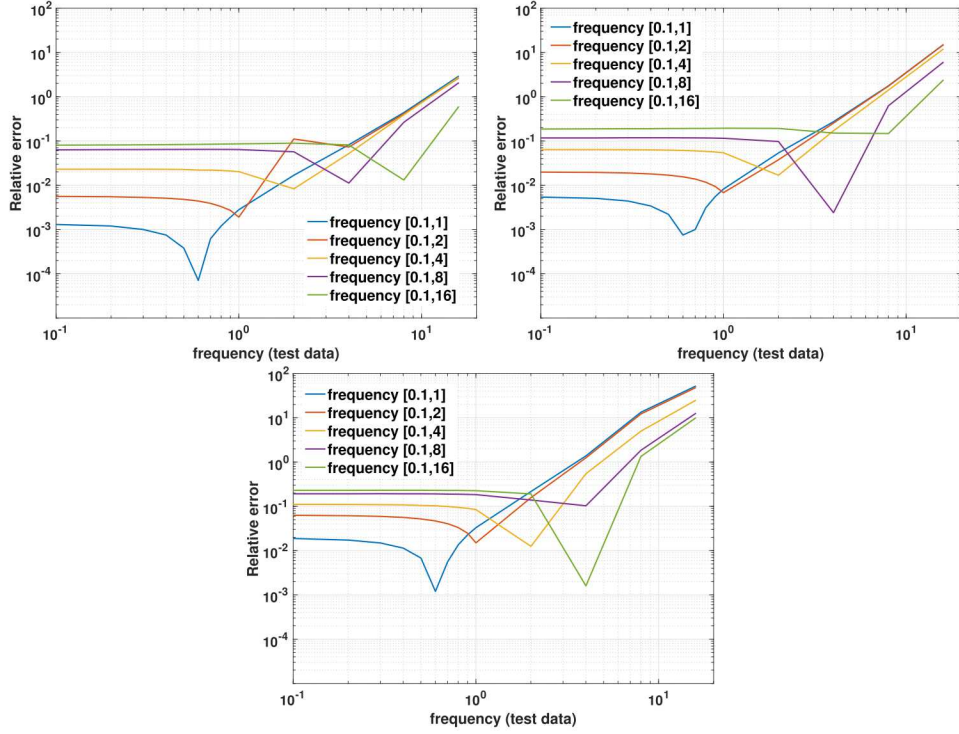
19

Figure 8: Relative solution error versus frequency of the forcing term in test data. The different curves in each subfigure signify a different range of frequency used in training data $[1/|\Omega|, 1/\lambda_{\min}]$. Increasing the ratio of $\lambda_{\min}/L$ provides improved fit. *Clockwise from top-left:* $\delta = 4L$, $\delta = 8L$, and $\delta = 16L$.
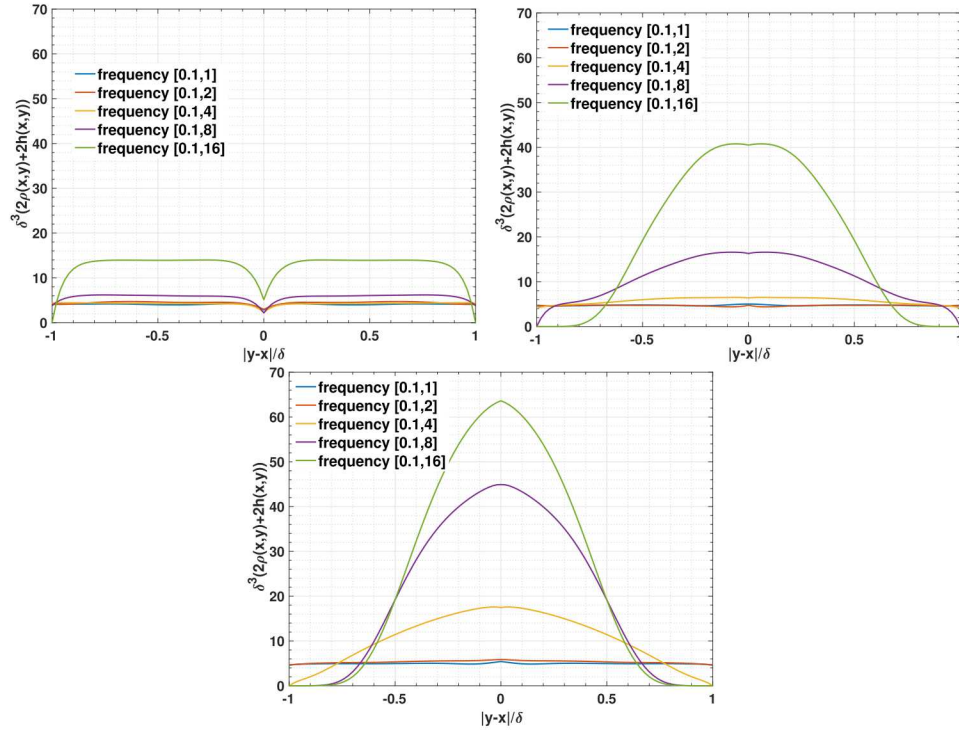

Figure 9: Trained kernels from different training data. Legend denotes range of frequency used in training data $[1/|\Omega|, 1/\lambda_{\min}]$. *Clockwise from top-left:* $\delta = 4L$, $\delta = 8L$, and $\delta = 16L$

20

[17, 47–51]. As an example, one may augment the description of a diffusion process by a Laplacian with higher-order operators such as biharmonics. Oftentimes these corrections provide a means to introduce lengthscales originating in microscales, e.g. Bazant introduces a correction of the form $c\delta^2\nabla^4$ to the Poisson-Boltzmann equation to correct for the effect of finite ion size [52, 53]. Such local corrections may be related to nonlocal models by assuming a large degree of regularity, expanding the solution in a Taylor series, and matching the moments of the kernel to corresponding local terms. However, nonlocal models possess the desirable feature of only requiring L^2-regularity. Weckner and Silling derived in this manner a sign-changing kernel, where the high-frequency response manifests as a negative part. We will consider a simplified problem, and illustrate that the introduction of a very small negative component allows for a substantial improvement in model accuracy.

We seek a nonlocal surrogate for the biharmonic equation

$$u_{xx} - c\delta^2 u_{xxxx} = f, \qquad (7.1)$$

u is periodic on $[0, L]$.

We generate data $\mathcal{D}_{\text{test}} = \{(u_i, f_i)\}_{i=1}^N$ for this equation for various c and δ by randomly generating

$$f_i = \sum_{k=1}^{99} \widehat{f}_k \cos(2\pi k x/L) \qquad (7.2)$$

for coefficients are sampled as

$$\widehat{f}_k \sim \exp(-\alpha k^2)\xi, \quad \xi \sim \mathcal{U}[0,1], \qquad (7.3)$$

where $\mathcal{U}[0,1]$ denotes the uniform distribution on $[0,1]$. For each sampled f_i, the corresponding u_i is solved from (7.1) using the discrete Fourier transformation. We generate $N = 50,000$ pairs of (u_i, f_i) for training, and train the kernel using an expansion (3.11) of order $M = 20$ for a range of δ from $\delta = 0.125$ to $\delta = 0.99$ and c from $c = 0.0001$ to $c = 0.1$.

In Table 1, we report the losses after the first stage of the algorithm, when $\mathbf{C}$ has been trained and $\mathbf{D} = 0$, and the loss after the second stage of the algorithm when both $\mathbf{C}$ and $\mathbf{D}$ have been trained. Figure 10 compares these two kernels, for $c = 0.0003$ and $\delta = 0.5$, to illustrate that the two kernels are similar except for a small negative tail in the sign-changing kernel for large $|y - x|/\delta$. Nevertheless, Table 1 illustrates that by fitting a sign-changing kernel as opposed to a nonnegative kernel, an average improvement in the loss by 18% for $\delta = 0.125$, 126% for $\delta = 0.25$, 98% for $\delta = 0.5$, and 12% for $\delta = 0.99$ is observed.

21

c	$\delta = 0.125$		$\delta = 0.25$		$\delta = 0.5$		$\delta = 0.99$	
	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$
0.0001	7.89e-5	6.49e-5	5.94e-4	1.48e-4	4.50e-3	1.50e-3	1.99e-2	1.78e-2
0.0003	8.47e-5	8.45e-5	6.08e-4	1.46e-4	4.87e-3	1.88e-3	2.27e-2	1.71e-2
0.0005	9.46e-5	8.30e-5	6.79e-4	1.99e-4	5.58e-3	2.44e-3	2.46e-2	1.81e-2
0.001	9.97e-5	7.77e-5	7.17e-4	3.11e-4	6.12e-3	2.07e-3	2.88e-2	2.45-02
0.003	1.46e-4	1.32e-4	1.10e-3	5.14e-4	1.11e-2	4.76e-3	4.84e-2	3.90e-2
0.005	1.95e-4	1.74e-5	1.73e-3	1.32e-3	1.52e-2	9.16e-3	6.11e-2	5.46e-2
0.01	3.52e-4	3.14e-4	3.18e-3	1.94e-3	2.20e-2	1.64e-2	8.16e-2	7.20e-2
0.03	1.30e-3	1.25e-3	1.04e-2	8.09e-3	5.09e-2	4.32e-2	1.04e-1	1.08e-1
0.05	2.61e-3	2.38e-3	1.91e-2	1.51e-2	6.77e-2	5.48e-2	1.10e-1	1.18e-1
0.1	6.35e-3	6.18e-3	3.37e-2	3.47e-2	8.81e-2	7.83e-2	1.34e-1	1.23e-1

Table 1: Training loss when using $\{C_m\}$ only and $\{C_m + D_m\}$ under different settings. The loss averaged over all c improves by 18%, 126%, 98%, and 12% for $\delta = 0.125, 0.25, 0.5$ and 0.99, respectively, when allowing for sign-changing kernels.
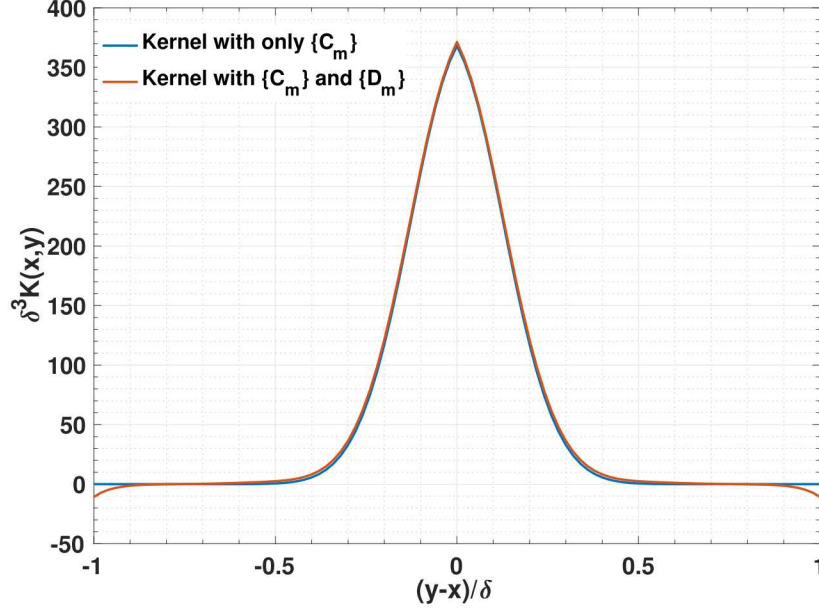

Figure 10: The comparison of the positive part of K^* determined by $\{C_m\}$ and the full kernel K^* given by both $\{C_m\}$ and $\{D_m\}$, when $c = 0.003$, $\delta = 0.5$. Note that allowing non-positive kernels provides only a very small negative part near $|y - x| = \delta$, yet provides a substantial increase in accuracy (Tables 1 and 2).

Table 2 compares the relative error in solving (7.1) for the test forcing

$$f = -4\pi^2 \sin(2\pi x) - 16c\delta^2\pi^4 \sin(2\pi x) \tag{7.4}$$

using the nonlocal model (3.11) with trained kernel K^*. We solve the nonlocal equation $\mathcal{L}_{K^*} u = f$ with periodic boundary condition. We compare the numerical solutions to the exact solution $u = \sin(2\pi x)$. This table corroborates Table 1, showing a significant in decrease testing error for $\delta = 0.25$ and $\delta = 0.5$ when using the sign-changing kernel rather than a strictly nonnegative kernel.

c	$\delta = 0.125$		$\delta = 0.25$		$\delta = 0.5$		$\delta = 0.99$	
	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$	$\{C_m\}$	$\{C_m + D_m\}$
0.0001	0.49%	0.39%	1.31%	0.19%	3.15%	0.09%	5.74%	1.37%
0.0003	0.48%	0.48%	1.27%	0.20%	3.39%	0.13%	5.66%	0.98%
0.0005	0.51%	0.54%	1.37%	0.25%	3.39%	0.13%	6.15%	1.42%
0.001	0.55%	0.44%	1.48%	0.46%	3.67%	0.12%	6.31%	0.87%
0.003	0.67%	0.61%	1.82%	0.52%	4.54%	0.15%	6.26%	0.86%
0.005	0.73%	0.71%	2.16%	1.11%	5.17%	0.24%	6.36%	0.94%
0.01	0.97%	1.04%	2.96%	1.05%	6.13%	0.42%	6.25%	1.27%
0.03	1.91%	1.67%	4.85%	2.59%	6.88%	0.20%	4.00%	0.97%
0.05	2.74%	2.33%	5.86%	2.97%	6.81%	0.19%	3.68%	0.71%
0.1	4.03%	3.69%	6.84%	5.17%	5.97%	1.29%	2.87%	0.95%

Table 2: Relative solution error for the test case (7.4) when using $\{C_m\}$ only and $\{C_m + D_m\}$ under different settings. Note that while introduction of D_m provides notable increase in accuracy, the resulting kernels differ by only a small negative tail (see Figure 10).

8. Fractional Laplacian

In this section we consider a high-fidelity model substantially different from the ones considered in Sections 6 and 7: the Poisson problem for the fractional Laplacian [16]. Fractional order equations are nonlocal equations often used in subsurface modeling; they accurately represent complex multi-scale phenomena by incorporating long-range interactions into the model itself in the form of a fractional-exponent derivative, as opposed to the classical integer-exponent derivative. This allows to capture the complete spectrum of diffusion through a few scalar model parameters. Despite being known to be the model of choice for scientific and engineering applications such as subsurface flow and transport [16], their adoption is limited due to several technical challenges, computational cost being the most important. This is understood by noting that the support of the nonlocal kernel is infinite, as opposed to the kernels considered in this paper whose support is limited to a ball of radius δ. Formally,

the fractional Laplacian operator is defined as

$$(-\Delta)^s u = C_{d,s} \int_{\mathbb{R}^d} \frac{u(x) - u(y)}{|x - y|^{d+2s}} dy, \tag{8.1}$$

where $s \in (0,1)$ and $C_{d,s}$ is a constant that depends on the dimension d and the fractional order s. Here, the *positive* nonlocal kernel, with support on $\mathbb{R}^d$ is given by

$$K(x,y) = \frac{C_{d,s}}{|x - y|^{d+2s}}. \tag{8.2}$$

We point out that, the interaction domain associated with the fractional Laplacian equation corresponds to the case of infinite horizon $\delta = \infty$. Thus, the high-fidelity problem we consider is (2.1) with $\mathcal{L}_{\mathrm{HF}} = (-\Delta)^s$ and an exterior condition $u = g$ on $\Omega_I = \mathbb{R}^d \setminus \Omega$. In other words, exterior conditions must be prescribed on an infinite domain.

Our goal in this section is to answer the following question: can we learn a nonlocal compactly supported kernel (with a relatively small horizon) from high-fidelity data from a fractional model such that the action of the corresponding nonlocal operator is equivalent to the fractional Laplacian in (8.1)? A positive answer would enable accurate computation of fractional solutions at a much cheaper cost and increase the usability of fractional models now hindered by computational challenges.

To answer the question above, we first note that solutions to the Poisson problem for the *truncated* fractional Laplacian (obtained by simply truncating the domain of integration in (8.1)) approach solutions of the non-truncated counterpart asymptotically with order $\mathcal{O}(\delta^{-2s})$ [54]; this type of convergence holds in general for other instances of fractional operators such as the fractional gradient and divergence [55]. Thus, for very small horizons, e.g., for $\delta \approx |\Omega|$, we cannot expect solutions to the truncated equation to be accurate approximations of solutions to the non-truncated one. This fact can be observed in the last column of Table 3 where we report the relative difference, for different values of δ between the non-truncated and truncated solutions; as expected, for very small δ, the truncated solution is significantly different than the non-truncated one. These results are reported as a reference and their importance will be clear after we describe the impact of our algorithm.

In order to reproduce the singularity in (8.2) we slightly modify the kernel expansion (3.11) by dividing each basis function $B_{m,M}$ by $|x - y|^\alpha$ instead of δ^{d+2}, where α is an additional tunable parameter. This means that we augment the set of parameters to be optimized within our algorithm by adding α. Furthermore, due to the positivity of the fractional kernel (8.2), we only perform the first stage of Algorithm 1 to learn the coefficients $\mathbf{C}$ and the parameter α. Recall that, as mentioned in the introduction, fractional equations

are master equations of stochastic processes and, as such, their kernels represent (positive) jump rates.

We generate the training set $\mathcal{D}_{\text{train}} = \{(u_i, f_i)\}_{i=1}^N$ by computing solutions of the fractional Laplacian equation with $s = 0.75$ in $\Omega = B_1(0)$, a ball of radius 1 and centered at 0, and exterior condition $g \equiv 0$ in $\mathbb{R}^d \setminus \Omega$. in We sample 50,000 forcing terms f_i as in (7.2) and (7.3), and obtain the corresponding solutions u_i by evaluating the integral

$$u_i(x) = \int_{B_1(0)} \mathcal{G}(x, y) f_i(y) dy,$$

using the trapezoidal quadrature rule, where $\mathcal{G}$ is the Green's function [54]. We test our algorithm for several values of δ and various order M of the basis, initializing α to be 0. In Table 3 we report the relative difference between the test solution for forcing $f = 1$ and fractional order $s = 0.75$ and the solution obtained by solving the nonlocal problem (2.2) with kernel K^*. While a higher order does not necessarily yield a better prediction, for higher values of δ, as expected, we obtain a much better approximation. As a reference we also report the difference that one would obtain by simply truncating the kernel (8.2). These result show that it is possible to obtain relatively accurate solutions to fractional equations at a much cheaper cost by using compactly supported fractional-type kernels learned from fractional data.

δ	relative difference				
	order $= 0$	order $= 5$	order $= 10$	order $=20$	truncated kernel
2^{-3}	9.46%	7.93%	16.8%	19.8%	319%
2^{-2}	15.4%	22.89%	23.5%	26.0%	182%
2^{-1}	13.2%	8.57%	10.0 %	7.2%	98%
2^0	3.97%	1.76%	2.37%	2.88%	48%
2^1	0.84%	2.18%	1.84%	3.43%	23%
2^2	1.59%	4.25%	3.66%	1.95%	15%

Table 3: Relative difference between the analytic solution with forcing $f = 1$ and fractional order $s = 0.75$ and the solution obtained in correspondence of the reconstructed kernel for several values of δ and different basis orders.

9. Conclusion

In this work we have presented an optimization framework for discovering sign-changing nonlocal models from high-fidelity synthetic data. In the nonlocal community there are several open questions about the derivation of such models from first-principles, and the data-driven approach presented here can provide guidance regarding what models emerge

naturally from data. The fundamental property we have pursued is a guarantee that learned models be solvable - this ensures robustness without requiring access to complicated PDE-constrained optimization codes, and may be implemented in popular optimization packages such as Tensorflow [56] and PyTorch [57]. While we worked with Bernstein polynomials for ease of implementing the resulting inequality constraints, an interesting area of future research would be to consider whether incorporating deep learning architectures into this framework provides benefit. Working with polynomials assumes regularity in the underlying kernels, while e.g. shallow ReLU networks would allow parameterization of discontinuous kernels [58, 59].

While we have focused on simple one-dimensional experiments for ease of presentation and to ensure examples are easily reproducible, the results demonstrate the potential impact this framework may have for range of important problems. We have demonstrated in Section 6 how one may use this approach to perform coarse-graining without incorporating restrictive and mathematically complex derivations. In Section 7, we illustrated the instrumental role sign-changing kernels have deriving reduced-regularity nonlocal versions of high-order PDEs, which have been shown previously to be fundamental in resolving high-frequency response of certain solid materials [17]. In Section 8, we see that one may extract computationally efficient compactly supported models providing sparse discrete operators that accurately approximate fractional operators with infinite horizons. This is particularly promising as a means of deriving preconditioners and $O(n)$ solvers for fractional systems, as naive discretizations provide dense matrices which require complex hierarchical preconditioners to solve efficiently [60, 61]. However, these examples provide only a first sample of possible directions to learn data-driven nonlocal models, and we pursue in future work application of this framework toward large-scale datasets more representative of open problems in science and engineering.

Acknowledgements

References

[1] Y. D. Ha, F. Bobaru, Characteristics of dynamic brittle fracture captured with peridynamics, Engineering Fracture Mechanics 78 (6) (2011) 1156–1168.

[2] D. J. Littlewood, Simulation of dynamic fracture using peridynamics, finite element modeling, and contact, in: ASME 2010 International Mechanical Engineering Congress and Exposition, American Society of Mechanical Engineers Digital Collection, 2010, pp. 209–217.

[3] H. Ouchi, A. Katiyar, J. York, J. T. Foster, M. M. Sharma, A fully coupled porous flow and geomechanics model for fluid driven cracks: a peridynamics approach, Computational Mechanics 55 (3) (2015) 561–576.

[4] P. Fife, Some nonclassical trends in parabolic and parabolic-like evolutions, in: Trends in nonlinear analysis, Springer, 2003, pp. 153–191.

[5] A. M. Tartakovsky, N. Trask, K. Pan, B. Jones, W. Pan, J. R. Williams, Smoothed particle hydrodynamics and its applications for multiphase flow and reactive transport in porous media, Computational Geosciences 20 (4) (2016) 807–834.

[6] R. Ansari, H. Rouhi, S. Sahmani, Calibration of the analytical nonlocal shell model for vibrations of double-walled carbon nanotubes with arbitrary boundary conditions using molecular dynamics, International Journal of Mechanical Sciences 53 (9) (2011) 786–792.

[7] P. Seleson, M. L. Parks, M. Gunzburger, R. Lehoucq, Peridynamics as an upscaling of molecular dynamics, Multiscale Modeling & Simulation 8 (1) (2009) 204–227.

[8] S. A. Silling, Reformulation of elasticity theory for discontinuities and long-range forces, Journal of the Mechanics and Physics of Solids 48 (1) (2000) 175–209.

[9] R. Arambakam, H. V. Tafreshi, B. Pourdeyhimi, Modeling performance of multi-component fibrous insulations against conductive and radiative heat transfer, International Journal of Heat and Mass Transfer 71 (2014) 341–348.

[10] A. V. Chebykin, A. A. Orlov, A. V. Vozianova, S. I. Maslovski, Y. S. Kivshar, P. A. Belov, Nonlocal effective medium model for multilayered metal-dielectric metamaterials, Physical Review B 84 (11) (2011) 115438.

[11] J. Xu, A. Askari, O. Weckner, S. Silling, Peridynamic analysis of impact damage in composite laminates, Journal of Aerospace Engineering 21 (3) (2008) 187–194.

[12] E. Askari, F. Bobaru, R. Lehoucq, M. L. Parks, S. A. Silling, O. Weckner, Peridynamics for multiscale materials modeling, in: Journal of Physics: Conference Series, Vol. 125, IOP Publishing, 2008, p. 012078.

[13] B. Alali, R. Lipton, Multiscale dynamics of heterogeneous media in the peridynamic formulation, Journal of Elasticity 106 (1) (2012) 71–103.

[14] A. Javili, R. Morasata, E. Oterkus, S. Oterkus, Peridynamics review, Mathematics and Mechanics of Solids 24 (11) (2019) 3714–3739.

[15] Q. Du, M. Gunzburger, R. Lehoucq, K. Zhou, Analysis and approximation of nonlocal diffusion problems with volume constraints, SIAM Review 54 (2012) 667–696.

[16] M. M. Meerschaert, J. Mortensen, S. W. Wheatcraft, Fractional vector calculus for fractional advection–dispersion, Physica A: Statistical Mechanics and its Applications 367 (2006) 181–190.

[17] O. Weckner, S. A. Silling, Determination of the constitutive model in peridynamics from experimental dispersion data, International Journal of Multiscale Computational Engineering, under review.

[18] A. H. Delgoshaie, D. W. Meyer, P. Jenny, H. A. Tchelepi, Non-local formulation for multiscale flow in porous media, Journal of Hydrology 531 (2015) 649–654.

[19] E. T. Chung, Y. Efendiev, W. T. Leung, M. Vasilyeva, Y. Wang, Non-local multicontinua upscaling for flows in heterogeneous fractured media, Journal of Computational Physics 372 (2018) 22–34.

[20] Q. Du, Z. Huang, R. Lehoucq, Nonlocal convection-diffusion volume-constrained problems and jump processes, Discrete & Continuous Dynamical Systems-Series B 19 (2).

[21] O. Defterli, M. D'Elia, Q. Du, M. Gunzburger, R. Lehoucq, M. M. Meerschaert, Fractional diffusion on bounded domains, Fractional Calculus and Applied Analysis 18 (2) (2015) 342–360.

[22] M. D'Elia, Q. Du, M. Gunzburger, R. Lehoucq, Nonlocal convection-diffusion problems on bounded domains and finite-range jump processes, Computational Methods in Applied Mathematics 29 (2017) 71–103.

[23] X. Xu, J. T. Foster, Deriving peridynamic influence functions for one-dimensional elastic materials with periodic microstructure, arXiv:2003.05520 (2020).

[24] T. Mengesha, Q. Du, Analysis of a scalar nonlocal peridynamic model with a sign changing kernel, Discrete & Continuous Dynamical Systems-B 18 (5) (2013) 1415–1437.

[25] M. D'Elia, M. Gunzburger, Identification of the diffusion parameter in nonlocal steady diffusion problems, Applied Mathematics & Optimization 73 (2) (2016) 227–249.

[26] G. Pang, M. D'Elia, M. Parks, G. E. Karniadakis, nPINNs: nonlocal Physics-Informed Neural Networks for a parametrized nonlocal universal Laplacian operator. Algorithms and Applications, arXiv:2004.04276 (2020).

[27] G. Pang, L. Lu, G. E. Karniadakis, fPINNs: Fractional physics-informed neural networks, SIAM Journal on Scientific Computing 41 (2019) A2603–A2626.

[28] M. D'Elia, C. Glusa, E. Otárola, A priori error estimates for the optimal control of the integral fractional laplacian, SIAM Journal on Control and Optimization 57 (4) (2019) 2775–2798.

[29] M. Gunzburger, R. Lehoucq, A nonlocal vector calculus with application to nonlocal boundary value problems, Multi. Model. Simul. 8 (2010) 1581–1620.

[30] Q. Du, M. Gunzburger, R. Lehoucq, K. Zhou, A nonlocal vector calculus, nonlocal volume-constrained problems, and nonlocal balance laws, Math. Mod. Meth. Appl. Sci. 23 (2013) 493–540.

[31] X. Chen, M. Gunzburger, Continuous and discontinuous finite element methods for a peridynamics model of mechanics, Computer Methods in Applied Mechanics and Engineering 200 (9-12) (2011) 1237–1250.

[32] S. Silling, E. Askari, A meshfree method based on the peridynamic model of solid mechanics, Computers and Structures 83 (2005) 1526–1535.

[33] N. Trask, H. You, Y. Yu, M. L. Parks, An asymptotically compatible meshfree quadrature rule for nonlocal problems with applications to peridynamics, Computer Methods in Applied Mechanics and Engineering 343 (2019) 151–165.

[34] H. You, X. Y. Lu, N. Trask, Y. Yu, An asymptotically compatible approach for Neumann-type boundary condition on nonlocal problems, arXiv:1908.03853 (2019).

[35] M. D'Elia, Q. Du, C. Glusa, M. Gunzburger, X. Tian, Z. Zhou, Numerical methods for nonlocal and fractional models, Acta NumericaTo appear.

[36] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv:1412.6980 (2014).

[37] Y. Yu, J. Chen, T. Gao, M. Yu, DA: Dag structure learning with graph neural networks, arXiv:1904.10098 (2019).

[38] J. Nocedal, S. Wright, Numerical Optimization, Springer Science & Business Media, 2006.

[39] H. P. G. Darcy, Les Fontaines publiques de la ville de Dijon. Exposition et application des principes à suivre et des formules à employer dans les questions de distribution d'eau, etc, V. Dalamont, 1856.

[40] S. Whitaker, Flow in porous media i: A theoretical derivation of darcy's law, Transport in porous media 1 (1) (1986) 3–25.

[41] J. Carrera, A. Alcolea, A. Medina, J. Hidalgo, L. J. Slooten, Inverse problem in hydrogeology, Hydrogeology journal 13 (1) (2005) 206–222.

[42] L. E. Payne, J. F. Rodrigues, B. Straughan, Effect of anisotropic permeability on Darcy's law, Mathematical methods in the applied sciences 24 (6) (2001) 427–438.

[43] J.-L. Auriault, Transport in porous media: upscaling by multiscale asymptotic expansions, in: Applied micromechanics of porous materials, Springer, 2005, pp. 3–56.

[44] T. Arbogast, Mixed multiscale methods for heterogeneous elliptic problems, in: Numerical analysis of multiscale problems, Springer, 2012, pp. 243–283.

[45] M. Sen, E. Ramos, A spatially non-local model for flow in porous media, Transport in porous media 92 (1) (2012) 29–39.

[46] X. Du, M. Ostoja-Starzewski, On the size of representative volume element for Darcy law in random media, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 462 (2074) (2006) 2949–2963.

[47] A. C. Eringen, Linear theory of nonlocal elasticity and dispersion of plane waves, International Journal of Engineering Science 10 (5) (1972) 425–435.

[48] A. C. Eringen, Vistas of nonlocal continuum physics, International Journal of Engineering Science 30 (10) (1992) 1551–1565.

[49] A. C. Eringen, D. G. B. Edelen, On nonlocal elasticity, International Journal of Engineering Science 10 (3) (1972) 233–248.

[50] K. Jakata, A. G. Every, Determination of the dispersive elastic constants of the cubic crystals ge, si, gaas, and insb, Physical Review B 77 (17) (2008) 174301.

[51] R. D. Mindlin, Second gradient of strain and surface-tension in linear elasticity, International Journal of Solids and Structures 1 (4) (1965) 417–438.

[52] M. Z. Bazant, B. D. Storey, A. A. Kornyshev, Double layer in ionic liquids: Overscreening versus crowding, Physical Review Letters 106 (4) (2011) 046102.

[53] B. D. Storey, M. Z. Bazant, Effects of electrostatic correlations on electrokinetic phenomena, Physical Review E 86 (5) (2012) 056303.

[54] M. D'Elia, M. Gunzburger, The fractional Laplacian operator on bounded domains as a special case of the nonlocal diffusion operator, Computers and Mathematics with applications 66 (2013) 1245–1260.

[55] M. D'Elia, M. Gulian, H. Olson, G. E. Karniadakis, A unified theory of fractional, nonlocal, and weighted nonlocal vector calculus, sandia Report SAND2020-4869, Sandia National Laboratories (2020).

[56] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org (2015).

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 8024–8035.

[58] E. C. Cyr, M. A. Gulian, R. G. Patel, M. Perego, N. A. Trask, Robust training and initialization of deep neural networks: An adaptive basis viewpoint, arXiv preprint arXiv:1912.04862.

[59] J. He, L. Li, J. Xu, C. Zheng, Relu deep neural networks and linear finite elements, arXiv preprint arXiv:1807.03973.

[60] M. Ainsworth, C. Glusa, Towards an efficient finite element method for the integral fractional Laplacian on polygonal domains, in: Contemporary Computational Mathematics-A Celebration of the 80th Birthday of Ian Sloan, Springer, 2018, pp. 17–57.

[61] K. Xu, E. Darve, Efficient numerical method for models driven by Lévy process via hierarchical matrices, arXiv preprint arXiv:1812.08324.