Data-adaptive RKHS regularization for learning kernels in operators

Fei Lu

Department of Mathematics, Johns Hopkins University

Numerical Analysis Seminar @UMD, March, 2024

Learning kernels in operators

Learn the kernel ϕ:

$$R_\phi[u] + \epsilon = f$$

from data:

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

- Operator $R_\phi[u](x) = \int \phi(x-y) g[u](x,y) dy$

Learning kernels in operators

Learn the kernel ϕ: $$R_\phi[u] + \epsilon = f$$

from data: $$\mathcal{D} = \{(u_k, f_k)\}_{k=1}^N, \quad (u_k, f_k) \in \mathbb{X} \times \mathbb{Y}$$

- Operator $R_\phi[u](x) = \int \phi(x-y) g[u](x,y) dy$
  - Interacting particles/agents

$$R_\phi[u] = \nabla \cdot [u(K_\phi * u)] = \partial_t u - \sigma \Delta u, \quad K_\phi(x) = \phi(|x|)\frac{x}{|x|} \in \mathbb{R}^d$$

$$R_\phi[\boldsymbol{X}_t] = \Big[-\frac{1}{n}\sum_{j=1}^n K_\phi(X_t^i - X_t^j)\Big]_i = \dot{\boldsymbol{X}}_t + \dot{\mathbf{W}}_t, \qquad \mathbb{R}^{nd}$$

  - Nonlocal PDEs:

$$R_\phi[u](x) = \int_\Omega \phi(x-y)[u(y) - u(x)]dy = \partial_{tt} u - v.$$

  - Integral operators, Toeplitz matrix: $R_\phi u = (\phi(x_i - x_j) u_j) = f$

Learning kernels in operators

Learn the kernel ϕ:

$$R_\phi[u] + \epsilon = f$$

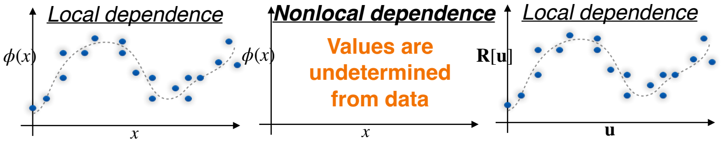
from data:

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

- Operator $R_\phi[u](x) = \int \phi(x-y) g[u](x,y) dy$
- Statistical learning $\bigcap$ inverse problem
  - random $\{(u_k, f_k)\}$: statistical learning
  - deterministic (e.g., N small): inverse problem

Learning kernels in operators

Classical learning	Learning kernels	Operator learning
$\{(x_i, \phi(x_i) + \epsilon_i)\}$	$\{(u_k, R_\phi[u_k] + \eta_k)\}$	$\{(u_k, R[u_k] + \eta_k)\}$



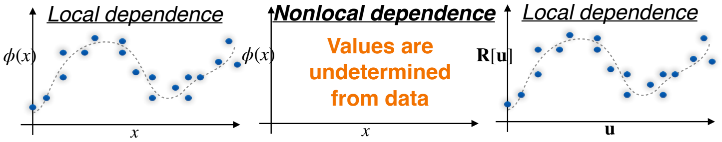
- **Nonlocal dependence**
- low-dimensional structure; linear in ϕ
- methods: regression/Neural network

Learning kernels in operators

Classical learning $\{(x_i, \phi(x_i) + \epsilon_i)\}$

Learning kernels $\{(u_k, R_\phi[u_k] + \eta_k)\}$

Operator learning $\{(u_k, R[u_k] + \eta_k)\}$



- **Nonlocal dependence**
- low-dimensional structure; linear in ϕ
- methods: regression/Neural network

This talk: $\Rightarrow$ Convergent estimator as mesh refines

- understand the **ill-posed** inverse problem
- introduce a new **regularization norm**

Part 2: Regression and regularization

Nonparametric regression

Loss functional: $\mathcal{E}(\phi) = \frac{1}{N}\sum_{i=1}^{N} \|R_\phi[u_i] - f_i\|_{L^2}^2.$

Hypothesis space: $\phi = \sum_{i=1}^{n} c_i \phi_i \in \mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^{n}$:

$$\mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^f, \Rightarrow \widehat{\phi}_{\mathcal{H}_n} = \sum_i \widehat{c}_i \phi_i, \text{ where } \widehat{c} = \overline{A}_n^{-1} \overline{b}_n,$$

Nonparametric regression

Loss functional: $\mathcal{E}(\phi) = \frac{1}{N}\sum_{i=1}^{N} \|R_\phi[u_i] - f_i\|_{L^2}^2.$

Hypothesis space: $\phi = \sum_{i=1}^{n} c_i \phi_i \in \mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^{n}$:

$$\mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^f, \Rightarrow \widehat{\phi}_{\mathcal{H}_n} = \sum_i \widehat{c}_i \phi_i, \text{ where } \widehat{c} = \overline{A}_n^{-1} \overline{b}_n,$$

Three issues

- $\overline{A}^{-1}$: ill-conditioned/singular
- Choice of $\mathcal{H}_n$: $\{\phi_i\}_{i=1}^{n}$ and n
- Convergence when data mesh refines $\Delta x \to 0$

Regularization

Regularization is necessary:

- $\overline{A}_n$ ill-conditioned
- $\overline{b}_n$: noise or numerical error

Tikhonov/ridge Regularization:

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|_*^2 \Rightarrow c^\top \overline{A}_n c - 2\overline{b}_n^\top c + \lambda \|c\|_{B_*}^2$$

$$\widehat{\phi}_{\mathcal{H}_n}^\lambda = \sum_i \widehat{c}_i^\lambda \phi_i, \quad \text{where } \widehat{c} = (\overline{A}_n + \lambda B_*)^{-1} \overline{b}_n,$$

Regularization

Regularization is necessary:

- $\overline{A}_n$ ill-conditioned
- $\overline{b}_n$: noise or numerical error

Tikhonov/ridge Regularization:

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|_*^2 \Rightarrow c^\top \overline{A}_n c - 2\overline{b}_n^\top c + \lambda \|c\|_{B_*}^2$$

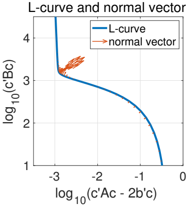
$$\widehat{\phi}_{\mathcal{H}_n}^\lambda = \sum_i \widehat{c}_i^\lambda \phi_i, \quad \text{where } \widehat{c} = (\overline{A}_n + \lambda B_*)^{-1} \overline{b}_n,$$

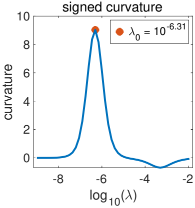
- λ by the L-curve method [Hansen00]

$$(x(\lambda), y(\lambda)) := (\log(\mathcal{E}(\widehat{c_\lambda})), \log(\|\widehat{c_\lambda}\|_*^2)),$$

$$\lambda_* = \text{maximal curvature}$$

- Which norm $\|\cdot\|_*$ to use? $B_* = I_n$?





Principle: [Stuart2010]

Avoid **discretization** until the last possible moment

$\downarrow$

Avoid basis selection until the last possible moment

Hypothesis space: $\phi = \sum_{i=1}^{n} c_i \phi_i \in \mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^{n}$:

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

Function space of ϕ? Identifiability?

Part 3: Identifiability & regularization

DARTR: Data adpative RKHS Tikhonov regularization

Identifiability

- An exploration measure: $\rho(dr) \qquad \Rightarrow \phi \in L^2_\rho$

  $R_\phi[u](x) = \int_\Omega \phi(|x-y|)g[u](x,y)dy, \quad \rho(dr) \propto \int\int \delta_{|x-y|}(dr)\big|g[u](x,y)\big|dxdy$

Identifiability

- An exploration measure: $\rho(dr) \qquad \Rightarrow \phi \in L^2_\rho$

  $R_\phi[u](x) = \int_\Omega \phi(|x-y|) g[u](x,y)dy, \quad \rho(dr) \propto \int\int \delta_{|x-y|}(dr)\big|g[u](x,y)\big|dxdy$

- An integral operator $\Leftarrow$ the Fréchet derivative of loss functional

$$\mathcal{E}(\psi) = \frac{1}{N}\sum_{i=1}^{N} \|R_\psi[u_i] - f_i\|^2_{L^2} = \langle \mathcal{L}_{\overline{G}}\psi, \psi\rangle_{L^2_\rho} - 2\langle \phi^D, \psi\rangle_{L^2_\rho}$$

$$\nabla\mathcal{E}(\psi) = 2\mathcal{L}_{\overline{G}}\psi - 2\phi^D = 0 \quad \Rightarrow \widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi^D$$

  - $\mathcal{L}_{\overline{G}}$: nonnegative compact, $\{(\lambda_i, \psi_i)\}$, $\lambda_i \downarrow 0$
  - $\phi^D = \mathcal{L}_{\overline{G}}\phi_{true} + \phi^{\mathrm{error}}$

Identifiability

- An exploration measure: $\rho(dr) \quad\quad \Rightarrow \phi \in L^2_\rho$

  $R_\phi[u](x) = \int_\Omega \phi(|x-y|) g[u](x,y)dy, \quad \rho(dr) \propto \int\int \delta_{|x-y|}(dr) |g[u](x,y)| dxdy$

- An integral operator $\Leftarrow$ the Fréchet derivative of loss functional

$$\mathcal{E}(\psi) = \frac{1}{N}\sum_{i=1}^{N} \|R_\psi[u_i] - f_i\|^2_{L^2} = \langle \mathcal{L}_{\overline{G}}\psi, \psi\rangle_{L^2_\rho} - 2\langle \phi^D, \psi\rangle_{L^2_\rho}$$

$$\nabla\mathcal{E}(\psi) = 2\mathcal{L}_{\overline{G}}\psi - 2\phi^D = 0 \quad\quad \Rightarrow \widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1}\phi^D$$

  - $\mathcal{L}_{\overline{G}}$: nonnegative compact, $\{(\lambda_i, \psi_i)\}$, $\lambda_i \downarrow 0$
  - $\phi^D = \mathcal{L}_{\overline{G}}\phi_{true} + \phi^{\text{error}}$

- Function space of identifiability (**FSOI**):

$$\widehat{\phi} = \mathcal{L}_{\overline{G}}^{-1}(\mathcal{L}_{\overline{G}}\phi_{true} + \phi^{\text{error}}) \Rightarrow \quad H = \mathrm{Null}(\mathcal{L}_{\overline{G}})^\perp = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}}$$

  - ill-defined beyond H; ill-posed in H

DARTR: Data Adaptive RKHS Tikhonov Regularization

A new task for Regularization:

ensure that the learning takes place in the FSOI

data-dependent $H = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}}$

DARTR: Data Adaptive RKHS Tikhonov Regularization

A new task for Regularization:

ensure that the learning takes place in the FSOI

$$\text{data-dependent} \quad H = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}} = \overline{H_G}^{L^2_\rho}$$

- $\overline{G} \Rightarrow$RKHS: $H_G = \mathcal{L}_{\overline{G}}^{1/2}(L^2_\rho)$
- For $\phi = \sum_k c_k\psi_k$, $\|\phi\|^2_{L^2_\rho} = \sum_k c_k^2$,

$$\|\phi\|^2_{H_G} = \sum_k \lambda_k^{-1} c_k^2 = \langle \mathcal{L}_{\overline{G}}^{-1}\psi, \psi\rangle_{L^2_\rho}$$

DARTR: Data Adaptive RKHS Tikhonov Regularization

A new task for Regularization:

ensure that the learning takes place in the FSOI

$$\text{data-dependent} \quad H = \overline{\mathrm{span}\{\psi_i\}_{i:\lambda_i>0}} = \overline{H_G}^{L^2_\rho}$$

- $\overline{G} \Rightarrow$RKHS: $H_G = \mathcal{L}_{\overline{G}}^{1/2}(L^2_\rho)$
- For $\phi = \sum_k c_k \psi_k$, $\|\phi\|^2_{L^2_\rho} = \sum_k c_k^2$,

$$\|\phi\|^2_{H_G} = \sum_k \lambda_k^{-1} c_k^2 = \langle \mathcal{L}_{\overline{G}}^{-1} \psi, \psi \rangle_{L^2_\rho}$$

$\Rightarrow$ Regularization norm: $\|\phi\|^2_{H_G}$

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|^2_{H_G} = \langle (\mathcal{L}_{\overline{G}} + \lambda \mathcal{L}_{\overline{G}}^{-1})\phi, \phi \rangle_{L^2_\rho} - 2\langle \phi^D, \phi \rangle_{L^2_\rho}$$

$$\widehat{\phi}_\lambda = (\mathcal{L}_{\overline{G}} + \lambda \mathcal{L}_{\overline{G}}^{-1})^{-1} \phi^D = (\mathcal{L}_{\overline{G}}^2 + \lambda I)^{-1} \mathcal{L}_{\overline{G}} \phi^D$$

What DARTR has done: remove error outside FSOI:
(Adaptive to data; regularize in FSOI)

- No regularization:

$$\widehat{\phi} = {\mathcal{L}_{\overline{G}}}^{-1}\phi^D = {\mathcal{L}_{\overline{G}}}^{-1}(\mathcal{L}_{\overline{G}}\phi_{true} + \phi_H^{\text{error}} + \phi_{H^\perp}^{\text{error}})$$

- DARTR: $\|\phi_{H^\perp}^{\text{error}}\|_{H_G}^2 = \infty$

$$(\mathcal{L}_{\overline{G}} + \lambda {\mathcal{L}_{\overline{G}}}^{-1})^{-1}\phi^D = (\mathcal{L}_{\overline{G}} + \lambda {\mathcal{L}_{\overline{G}}}^{-1})^{-1}(\mathcal{L}_{\overline{G}}\phi_{true} + \phi_H^{\text{error}})$$

- l^2 or L^2 regularizer: with $C = \sum \phi_i \otimes \phi_j$ or $C = I$

$$(\mathcal{L}_{\overline{G}} + \lambda C)^{-1}\phi^D = (\mathcal{L}_{\overline{G}} + \lambda C)^{-1}(\mathcal{L}_{\overline{G}}\phi_{true} + \phi_H^{\text{error}} + \phi_{H^\perp}^{\text{error}})$$

DARTR: computation

$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda\|\phi\|_{H_G}^2 \Rightarrow c^\top A_n c - 2b_n^\top c + \lambda\|c\|_{B_{rkhs}}^2$$

Input: A_n, b_n and $B_n = (\langle \phi_i, \phi_{j,\rangle}\ L^2_\rho)_{i,j}$.
Output: reguarized estimator

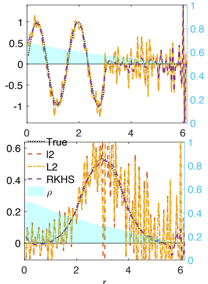
$$\widehat{c}_\lambda = (A_n + \lambda_* B_{rkhs})^{-1} b_n$$

- Generalized eigenvalue problem $(A_n, B_n) \leftrightarrow \mathcal{L}_{\overline{G}}$
  $A_n V = B_n V \Lambda$ and $V^\top B_n V = I_n$
  $B_{rkhs} = (V \Lambda V^\top)^\dagger$
- L-curve to select λ_*

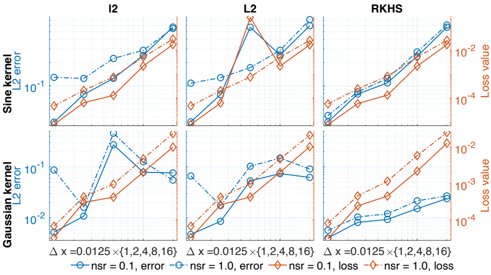
Interaction kernel in a nonlinear operator

$$R_{\phi}[u] = \nabla \cdot [u(K_{\phi} * u)] = f, \quad K_{\phi} = \phi(|x|)\frac{x}{|x|}$$

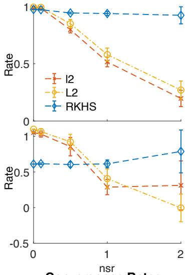
- Recover kernel from discrete noisy data
- Robust in accuracy, consistent rates as mesh refines



Typical estimators, $\Delta x = 0.05$

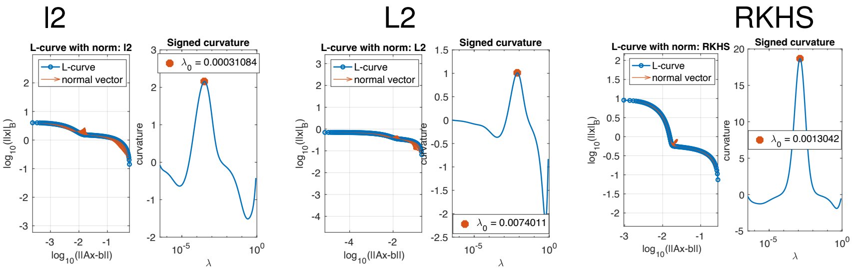


Convergence of Estimators, nsr = 0.1 & 1



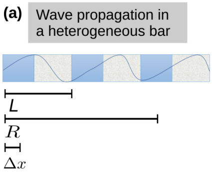
Convergence Rates

More robust L-curve



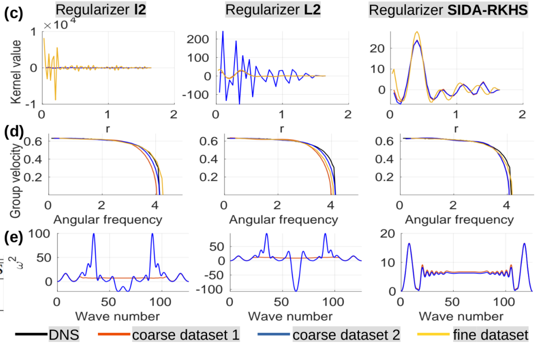
Homogenization of wave propagation in meta-material

- heterogeneous bar with microstructure + DNS $\Rightarrow$ Data
- Homogenization: [LAY23]
  $R_\phi[u] = \int_\Omega \phi(|y|)[u(x+y) - u(x)]dy = \partial_{tt} u - g.$



(b) Displacement error on a cross-validation dataset

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%**



- (c): resolution-invariant
- (e): l^2 and $L2$ leading to non-physical kernel

Part 4: Iterative method

Large scale $Ax = b$, $A \in \mathbb{R}^{m \times n}$ ill-conditioned , $n >> 1$

b: noisy

DARTR for $Ax = b$

$A_n = A^\top A, b_n = A^\top b$ and $B_n = \mathrm{diag}(\rho)$.

$$\widehat{c}_\lambda = (A_n + \lambda_* B_{rkhs})^{-1} b_n$$

- $\rho =$ normalized column sum of $(|A_{ij}|)$: pre-conditioning
- Generalized eigenvalue problem (A_n, B_n)
  $A_n V = B_n V \Lambda$ and $V^\top B_n V = I_n \Rightarrow B_{rkhs} = (V \Lambda V^\top)^\dagger$
  $B_{rkhs} = A_n^\dagger$ when $B_n = I_n$
- L-curve to select λ_*

DARTR for $Ax = b$

$A_n = A^\top A, b_n = A^\top b$ and $B_n = \mathrm{diag}(\rho)$.

$$\widehat{c}_\lambda = (A_n + \lambda_* B_{rkhs})^{-1} b_n$$

- $\rho =$ normalized column sum of $(|A_{ij}|)$: pre-conditioning
- Generalized eigenvalue problem (A_n, B_n)
  $A_n V = B_n V \Lambda$ and $V^\top B_n V = I_n \Rightarrow B_{rkhs} = (V \Lambda V^\top)^\dagger$
  $B_{rkhs} = A_n^\dagger$ when $B_n = I_n$
- L-curve to select λ_*

Direct method: based on **costly** matrix decomposition.

Iterative method: use but without computing B_{rkhs}?

Iterative Data Adaptive RKHS regularization

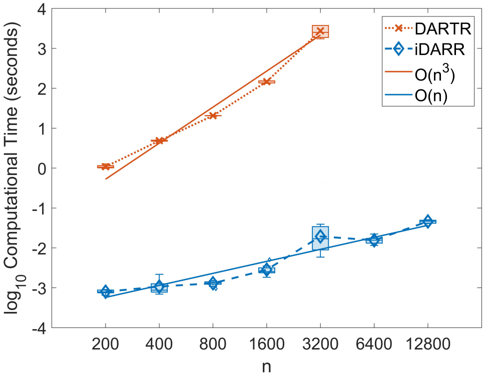
Solve: $x_k = \arg\min_{x \in \mathcal{X}_k} \|x\|_{B_{rkhs}}$, $\mathcal{X}_k = \{x : \min_{x \in \mathcal{S}_k} \|Ax - b\|\}$

$$\mathcal{S}_k = \mathrm{span}\{(B^{\dagger}_{rkhs} A^{\top} A)^i B^{\dagger}_{rkhs} A^{\top} b\}_{i=0}^{k}$$

- Use $B^{\dagger}_{rkhs}$, not B_{rkhs}: $B^{\dagger}_{rkhs} = B^{-1} A^{\top} A B^{-1}$
- generalized Golub-Kahan bidiagonalization (gGKB) $\Rightarrow$ construct $\mathcal{S}_k$ only using matrix-vector product
- $\mathcal{S}_k$ = RKHS-restricted Krylov subspace
- Early stopping: select k

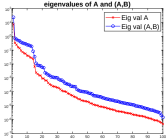
Computational complexity

DARTR: $O(n^3)$
iDARR: $O(3mnk)$



Fredholm integral equation: 1st kind

Polynomial decaying spectrum:



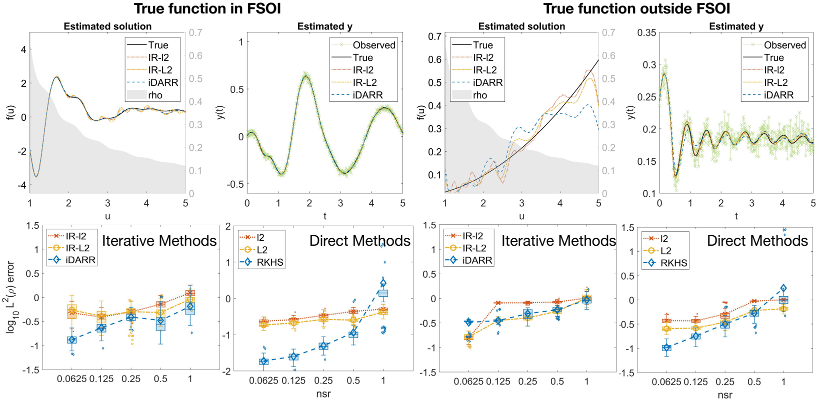
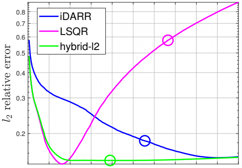


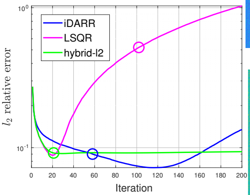
Image deblurring

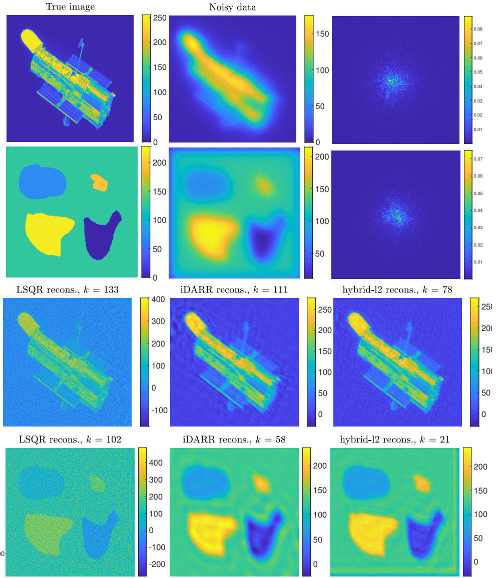
Image deblurring

Gazzola+Hansen+Nagy2019

256x256; 320x320







Regularization:

Is DA-RKHS better than other norms?

- No regularizer is universally "best"
  - no universal criteria: similar to Prior in Bayesian learning
  - Multiple factors: Smoothness of true function, Operator spectral decay, Noise distribution, hyper-parameter

Regularization:

Is DA-RKHS better than other norms?

- No regularizer is universally "best"
  - no universal criteria: similar to Prior in Bayesian learning
  - Multiple factors: Smoothness of true function, Operator spectral decay, Noise distribution, hyper-parameter
- Small noise analysis [CLLW22,LuOu23,LangLu23]
  - Data-Adaptive is important
    fractional RKHS $H_G^s = L_G^{s/2} L_\rho^2$
  - Convergence rate: same as L^2, a smaller factor
  - Robust for selection of hyper-parameter

Summary

Learning kernels in operators:

$$R_{\phi}[u] = f \quad \Leftarrow \quad \mathcal{D} = \{(u_k, f_k)\}_{k=1}^{N}$$

Nonlocal dependence

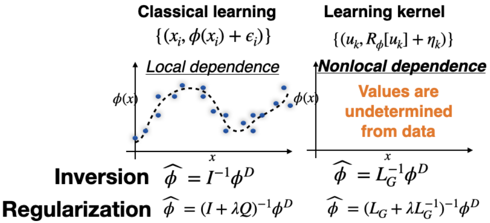
- Identifiability: FSOI
- DARTR: data adaptive RKHR Tikhonov-Reg
  - Synthetic data: convergent, robust to noise
  - Homogenization: resolution-independent
- Iterative method: iDARR

Regularization: $Ax = b \Rightarrow x_{\lambda} = (A + \lambda A^{-1})b$

Future directions

Learning with nonlocal dependence

- Convergence: $\Delta x, N$
- Data-adaptive basis
- Regularization for ML: $\|\phi_\theta\|^2_{rkhs}$, not $\|\theta\|$



References
- LLA22: Lu, Lang, and An. MSML22. (Matlab code)
- LAY22: Lu, An and Yu. J. Peridynamics& Nonlocal Modeling, 2023
- CLLW22: Chada, Lang, Lu, and Wang. arXiv2212
- LO23: Lu and Ou. arXiv2303.
- LL23: Lang and Lu, arXiv2305
- LFL24: Li, Feng and Lu, arXiv2401. (Matlab code)