DARTR: Data Adaptive RKHS Tikhonov Regularization for learning kernels in operators

Fei Lu

Department of Mathematics, Johns Hopkins University

Joint with **Quanjun Lang, Qingci An**@JHU

August, 2022; MSML

Learning kernels in operators

Learn the kernel ϕ:

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

from data:

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

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_ϕ: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles) [LangLu22sisc]

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

  - nonlocal PDE/ fractional diffusion :

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

  - Integral operators, deconvolution, Toeplitz/Hankel matrix ...

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_ϕ: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles) [LangLu22sisc]

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

  - nonlocal PDE/ fractional diffusion :

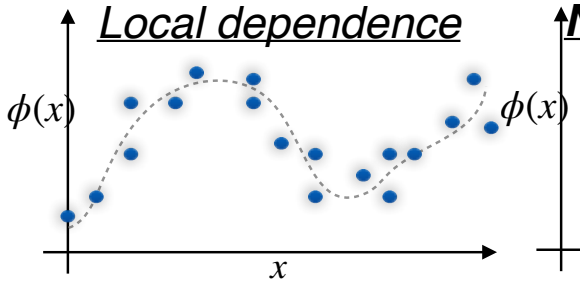
$$R_\phi[u](x) = \int_\Omega \phi(x,y)[u(y) - u(x)]dy = \partial_{tt} u - g, \forall x \in \Omega.$$

  - Integral operators, deconvolution, Toeplitz/Hankel matrix ...
- Data: discrete/noisy, **Nonlocal dependence**
  - random $(u_k, f_k) \sim \mu \otimes \nu$: statistical learning
  - deterministic (e.g., N=1): inverse problem

Comparison with classical learning

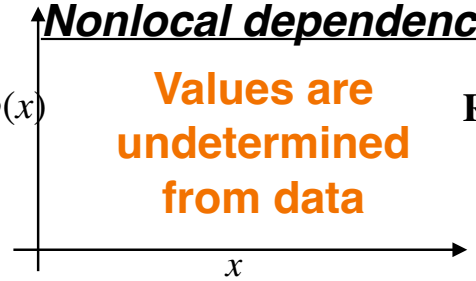
Classical learning

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



Learning kernels

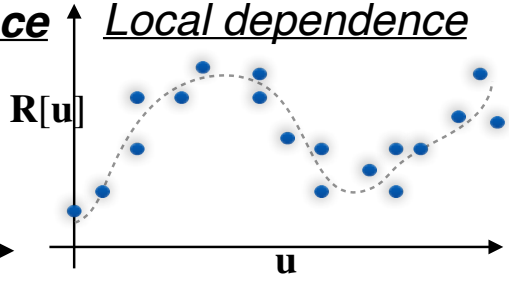
$\{(u_k, R_\phi[u_k] + \eta_k)\}$



Values are undetermined from data

Operator learning

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



- nonlocal dependence
  under-determined (no longer "interpolation")
- v.s. operator learning: low-dimensional structure
- methods: regression/ML/DL?

This talk: deterministic inverse problems

$$\mathcal{D} = \{u_k, f_k\}_{k=1}^N = \{u_k(x_j, t_l), f_k(x_j, t_l) : j = 1, \ldots, J\}_{i=1}^N,$$

Nonparametric regression

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

- Neural network when ϕ is high-D

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

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

Three issues

- $\overline{A}^{-1}$: well-posedness, Identifiability, function space
- 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/singular
- $\overline{b}_n$: 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]
- Which norm $\|\cdot\|_*$ to use?

Identifiability

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

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

Identifiability

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

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

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

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

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

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

Identifiability

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

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

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

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

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

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

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

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

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

DARTR: Data Adaptive RKHS Tikhonov Regularization

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

A new task for Regularization:

ensure that the learning takes place in the FSOI

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

DARTR: Data Adaptive RKHS Tikhonov Regularization

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

A new task for Regularization:
ensure that the learning takes place in the FSOI

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

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

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

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

Why DARTR is good: FSOI is fundamental:

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

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

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

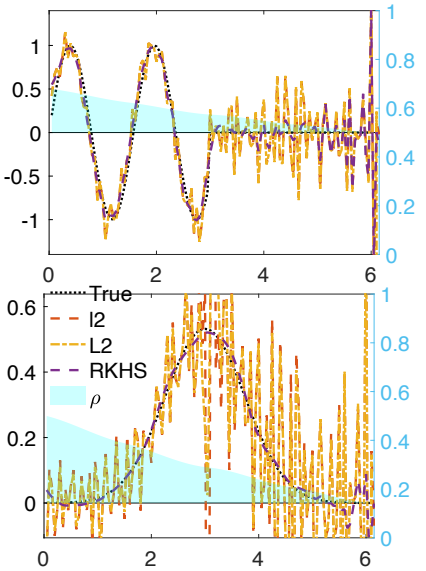
- l^2 or L^2 regularizer: with $C = \sum \phi_i \otimes \phi_j$ or $C = I$

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

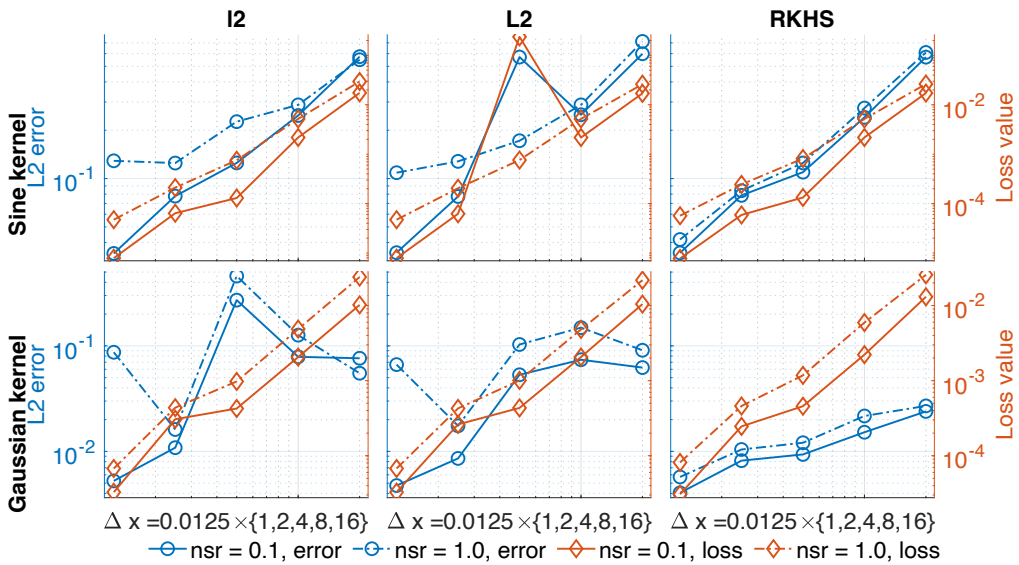
Interaction kernel in a nonlinear operator

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

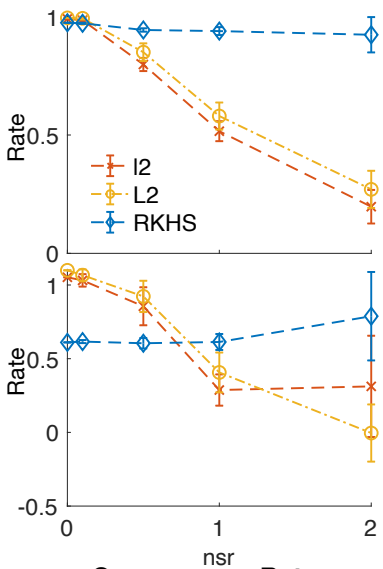
- Recover kernel from discrete noisy data
- Robust in accuracy, consistent rates as mesh refines



Typical estimators, $\Delta x = 0.05$



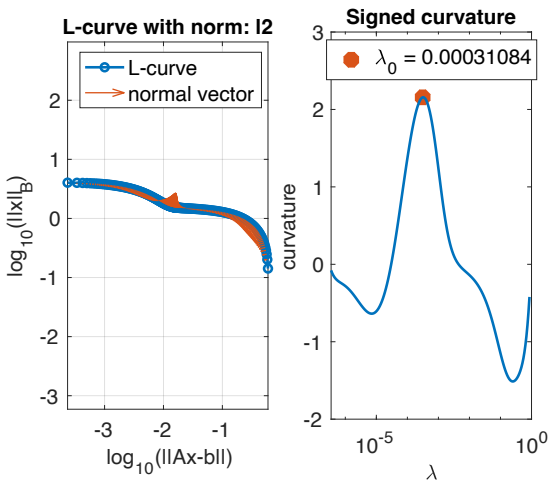
Convergence of Estimators, nsr = 0.1 & 1



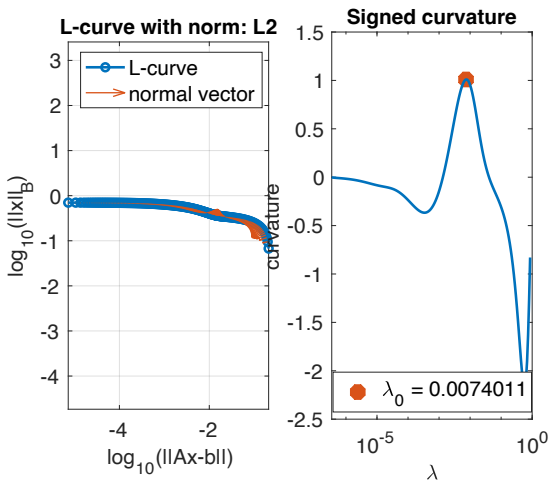
Convergence Rates

More robust L-curve

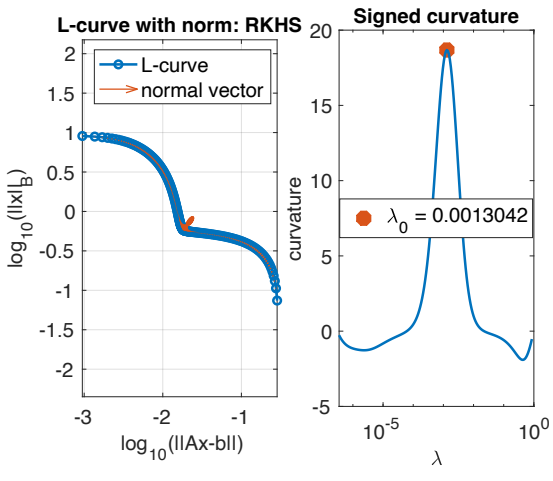
l2



L2

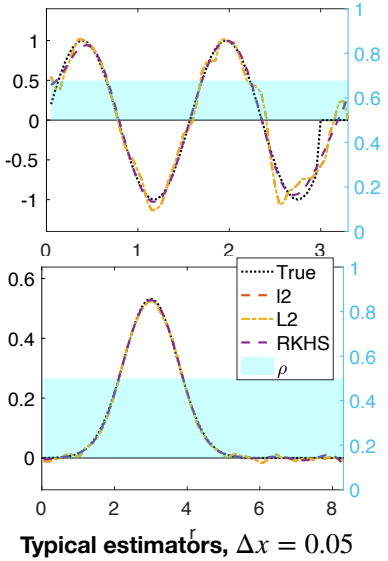


RKHS

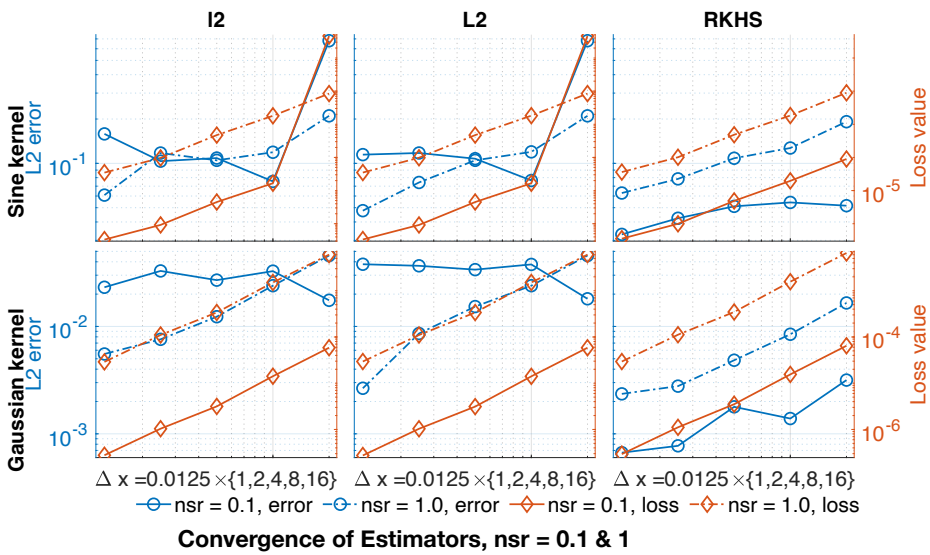


Linear integral operator:

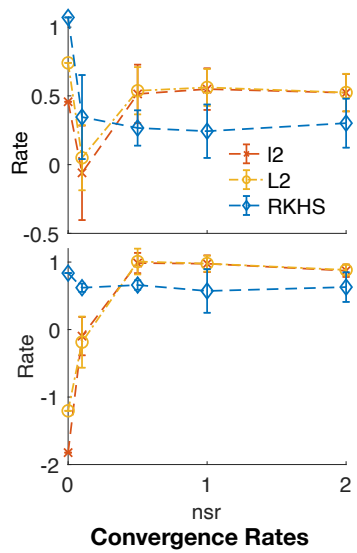
$$R_\phi[u](x) = \int_\Omega \phi(|y-x|)u(y)dy = f(x).$$



Typical estimators, $\Delta x = 0.05$



Convergence of Estimators, nsr = 0.1 & 1

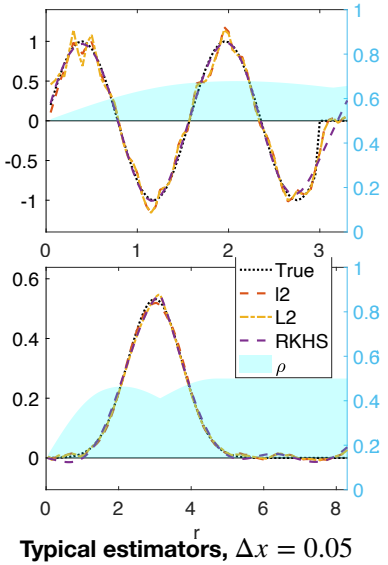


Convergence Rates

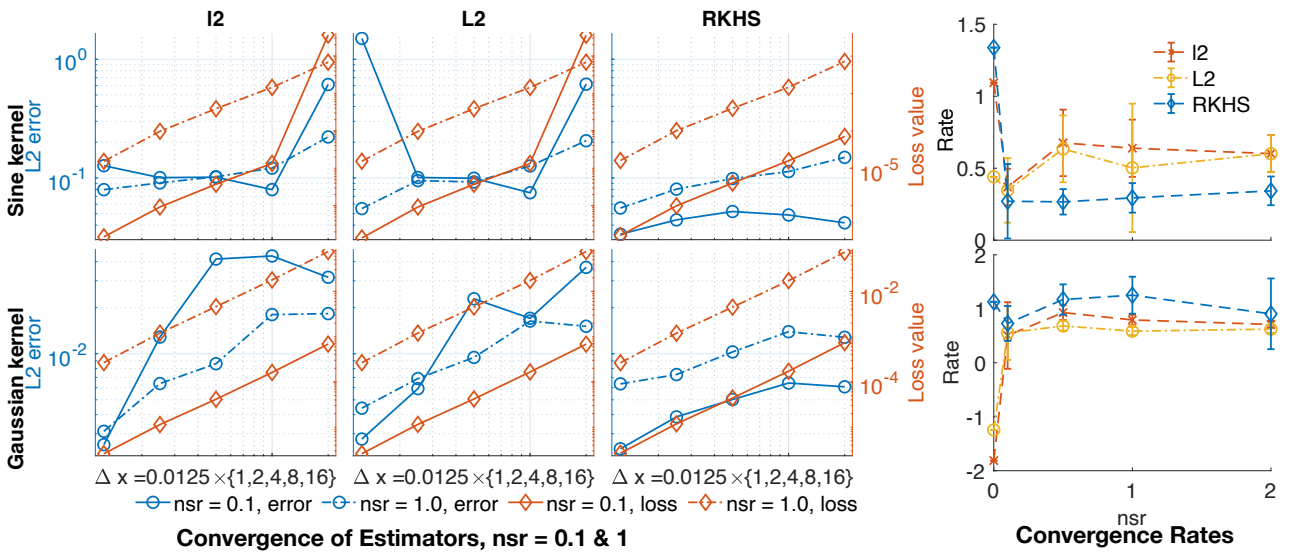
Robust in accuracy, consistent rates

Nonlocal operator

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



Typical estimators, $\Delta x = 0.05$



Convergence of Estimators, nsr = 0.1 & 1

Convergence Rates

Robust in accuracy, consistent rates

Summary

Learning kernels in operators:

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

Nonlocal dependence

- Identifiability theory: FSOI
- DARTR: data adaptive RKHR Tik-Reg
- Numerical tests: robust accuracy, consistent rates

Future directions

- Convergence of regularized estimator $(\Delta x, N)$
- Applications to linear inverse problems
- Regularization for neural network: $\|\phi_\theta\|^2_{rkhs}$

	Data	Goal: ϕ	Inversion*	FSOI	Regularization
Classical learning	$\{(x_i, y_i)\}$	$Y = \phi(X)$	$\widehat{\phi} = I^{-1}\phi^D$	$L^2(\rho)$	no need of FSOI
Learning kernels	$\{(u_i, f_i)\}$	$R_\phi[u] = f$	$\widehat{\phi} = \mathcal{L}_G^{-1}\phi^D$	SIDA	FSOI necessary