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, **Yue Yu**@Lehigh

MiC Seminars, ORNL; August, 2022

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 $R\phi$: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles, mean-field)

$$R_\phi[u] = \nabla\cdot[u(K_\phi * u)] = \partial_t u - \sigma\Delta u, \quad K_\phi(x) = \phi(|x|)\frac{x}{|x|} \in \mathbb{R}^d$$
$$R_\phi[\boldsymbol{X}_t] = \Big(-\frac{1}{n}\sum_{j=1}^n K_\phi(X_t^i - X_t^j)\Big)_i = \dot{\boldsymbol{X}}_t + \dot{\mathbf{W}}_t, \qquad \mathbb{R}^{nd}$$

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 $R\phi$: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles, mean-field)

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

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

  - nonlocal PDE/ fractional diffusion: $R_\phi[u] = \partial_{tt} u - g$

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

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 $R\phi$: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles, mean-field)

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

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

  - nonlocal PDE/ fractional diffusion: $R_\phi[u] = \partial_{tt} u - g$

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

  - Integral operators, deconvolution, Toeplitz/Hankel matrix ...
    Toeplitz matrix: $R_\phi u = f$, $R_\phi(i,j) = \phi(i-j)$

Learning kernels in operators

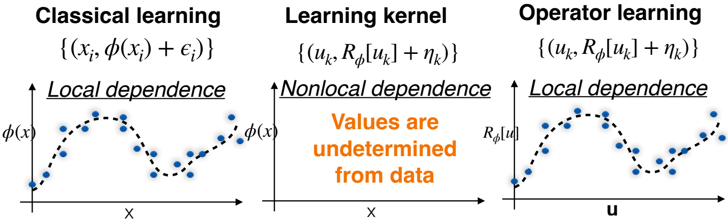
Learn the kernel ϕ: $$R_\phi[u] + \epsilon = f$$

from data:

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

- Operator $R\phi$: **linear or nonlinear in u, but linear in ϕ**
  - nonlocal interaction (interacting particles, mean-field)
  - nonlocal PDE/ fractional diffusion
  - 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



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

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

Three issues

- $\overline{A}^{-1}$: 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:

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

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

Regularization

Regularization is necessary:

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

Tikhonov/ridge Regularization:

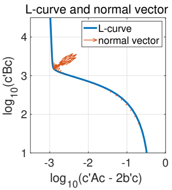
$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|_*^2 \Rightarrow c^\top \overline{A}_n c - 2\overline{b}_n^\top c + \lambda \|c\|_{B_*}^2$$

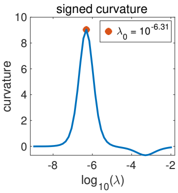
$$\widehat{\phi}_{\mathcal{H}_n}^\lambda = \sum_i \widehat{c}_i^\lambda \phi_i, \quad \text{where } \widehat{c} = (\overline{A}_n + \lambda B_*)^{-1} \overline{b}_n,$$

- λ by the L-curve method [Hansen00]

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

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





Regularization

Regularization is necessary:

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

Tikhonov/ridge Regularization:

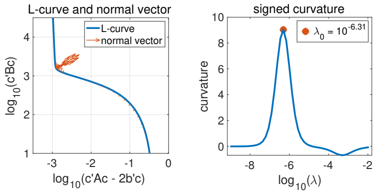
$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda \|\phi\|_*^2 \Rightarrow c^\top \overline{A}_n c - 2\overline{b}_n^\top c + \lambda \|c\|_{B_*}^2$$

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

- λ by the L-curve method [Hansen00]

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

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



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

Identifiability

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

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

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

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

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

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

Identifiability

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

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

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

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

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

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

DARTR: Data Adaptive RKHS Tikhonov Regularization

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

A new task for Regularization:

ensure that the learning takes place in the FSOI

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

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

- $\overline{G} \Rightarrow$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:

H

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

- 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 $C = I$

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

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 $C = I$

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

A Bayesian perspective:

- Prior $\mathcal{N}(0, \mathcal{L}_{\overline{G}})$; v.s. $\mathcal{N}(0, C)$: singular or equivalent
- Posterior $\mathcal{N}(\widehat{\phi}_*, (\mathcal{L}_{\overline{G}} + \lambda{\mathcal{L}_{\overline{G}}}^{-1})^{-1})$ v.s. $\mathcal{N}(\widehat{\phi}_\dagger, (\mathcal{L}_{\overline{G}} + \lambda C)^{-1})$
- Zellner's g-prior $\mathcal{N}(0, \overline{A}_n^{-1})$ if $\mathcal{H}_n = \mathrm{span}\{\phi_i\}_{i=1}^n$ o.n.b.

DARTR: compute the SIDA-RKHS norm

Let $B_n = (\langle \phi_i, \phi_j \rangle_{L^2(\rho)})_{i,j}$.

Theorem (Generalized eigenvalue problem)

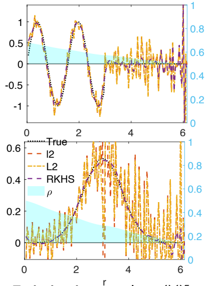
If $\mathcal{L}_{\overline{G}}(L^2(\rho)) \subset \mathcal{H}$, then $\mathcal{L}_{\overline{G}}$ eigenvalues are solved by the generalized eigenvalue problem $(\overline{A}_n, B_n)$ and $B_{rkhs} = (V \Lambda V^\top)^{-1}$.

- If $B_n = I_n$: $B_{rkhs} = \overline{A}_n^{-1}$, the Zellner's g-prior;
- If $\phi_i = \psi_i$: $\overline{A}_n = \mathrm{diag}(\lambda_i)$, $B_{rkhs} = \overline{A}_n^{-1}$:
  $\widehat{c} = \sum_{i:\lambda_i>0} (\lambda_i + \lambda)^{-1} (v_i^\top \overline{b}) v_i$
- l^2 or L^2 regularizer: $\widehat{c} = [\sum_{i:\lambda_i>0} + \sum_{i:\lambda_i=0}] (\lambda_i + \lambda)^{-1} (v_i^\top \overline{b}) v_i$

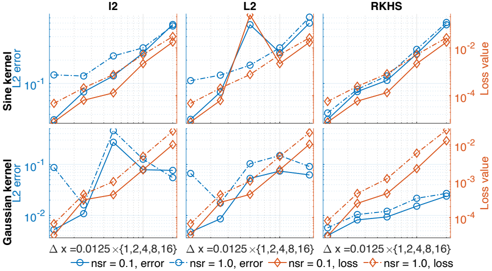
Interaction kernel in a nonlinear operator

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

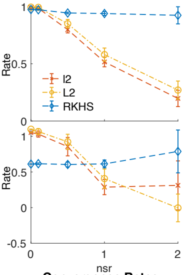
- Recover kernel from discrete noisy data
- Robust in accuracy, consistent rates as mesh refines



Typical estimators, $\Delta x = 0.05$



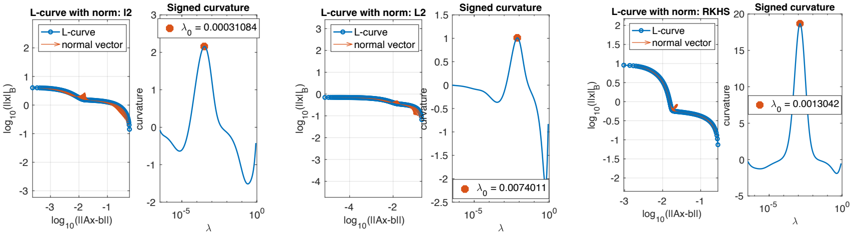
Convergence of Estimators, nsr = 0.1 & 1



Convergence Rates

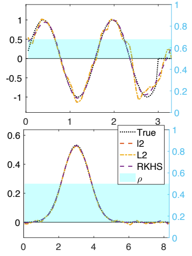
More robust L-curve

l2 · L2 · RKHS

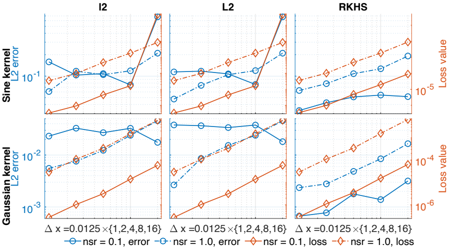


Linear integral operator:

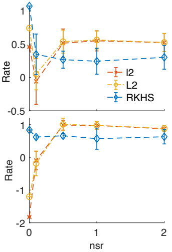
$$R_\phi[u](x) = \int_\Omega \phi(|y-x|)u(y)dy = f(x).$$



Typical estimators, $\Delta x = 0.05$



Convergence of Estimators, nsr = 0.1 & 1



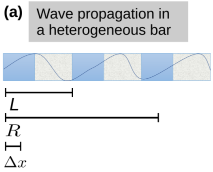
Convergence Rates

Robust in accuracy, consistent rates

Homogenization of wave propagation in meta-material

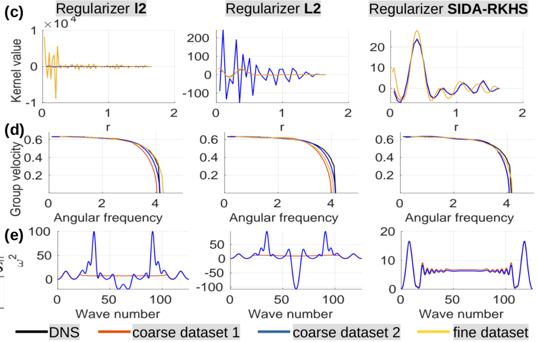
- heterogeneous bar with microstructure + DNS $\Rightarrow$ Data
- Homogenization: $R_\phi[u] = \partial_{tt} u - g$.

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



(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

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 Tikhonov-Reg
- Numerical tests: robust accuracy, consistent rates

Future directions

- Convergence of regularized estimator $(\Delta x, N)$
- Inverse problems with nonlocal dependence
- Regularization for neural network: $\|\phi_\theta\|^2_{rkhs}$, not $\|\theta\|$

	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

References

(@ `http://www.math.jhu.edu/~feilu`) QR-code →

- F.Lu, Q .Lang and Q. An. Data adaptive RKHS Tikhonov regularization for learning kernels in operators. MSML22. (Matlab code available)
- F.Lu, Q .An and Y. Yu. Nonparametric learning of kernels in nonlocal operators. arXiv2205