# Introduction to Nonparametric Learning of Kernels in Operators

Fei Lu

Department of Mathematics, Johns Hopkins University

Plan:
Lecture 1. Overview and a review of classical learning theory
Lecture 2. Learning interaction kernels in interacting particle systems
Lecture 3. Coercivity condition and minimax rate of convergence
Lecture 4. Learning interaction kernels in mean-field equations
Lecture 5. Data adaptive RKHS Tikhonov regularization
Lecture 6. Small noise analysis of RKHS regularizations

# Lec6. Small noise analysis for Tikhonov regularization

Learn the kernel $\phi$:

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

from data:

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

Variational approach: ill-posed $\Rightarrow$ **Regularization**, Tikhonov

$$\widehat{\phi}_\lambda = \arg\min_{\phi \in \mathcal{H}} \mathcal{E}(\phi) + \lambda \|\phi\|_*^2$$

- Regression:
  $\phi = \sum_{k=1}^n c_k \phi_k, \qquad A_n c = b_n$

  $$\|A_n c - b_n\|^2 + \lambda \|c\|_*^2$$

- Regularization norms:
  $l^2$, $L^2$, RKHSs, $H^1$ ...

**Which norm is better? Proof?**

1. Review: learning kernels
2. Why is DARTR good?
3. SNA for DARTR
4. SNA for fractional DARTR

- LO23: Lu+Ou arXiv 2303
- LL23: Lang+Lu arXiv2305.

# Outline

1. Review: learning kernels
2. Why is DARTR good?
3. SNA for DARTR
4. SNA for fractional DARTR

# Learning kernels in operators

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

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

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

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

  - nonlocal PDEs: $R_\phi[u] = \partial_{tt} u - v$

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

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

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

- Operator $R_\phi[u](x) = \int \phi(|x-y|) g[u](x,y) dy$
- $R_\phi[u]$ **linear in** $\phi$
- Data: discrete/noisy, **Nonlocal dependence**
  - random $(u_k, f_k) \sim \mu \otimes \nu$: statistical learning
  - deterministic (e.g., N small): inverse problem

# Learning kernels in operators

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

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

- Operator $R_\phi[u](x) = \int \phi(|x-y|) g[u](x,y) dy$
- $R_\phi[u]$ **linear in** $\phi$
- Data: discrete/noisy, **Nonlocal dependence**
  - random $(u_k, f_k) \sim \mu \otimes \nu$: statistical learning
  - deterministic (e.g., N small): inverse problem

Nonparametric inference $\Leftrightarrow$ Variational inverse problem

$$\widehat{\phi} = \arg\min_{\phi \in \mathcal{H}} \mathcal{E}(\phi), \quad \mathcal{E}(\phi) = \frac{1}{N} \sum_{i=1}^N \|R_\phi[u_i] - f_i\|_{\mathbb{Y}}^2.$$

# Computation: Regression and Regularization

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

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

**Regularization** necessary: $\overline{A}_n$ ill-conditioned & $\overline{b}_n$: noisy or with error
Tikhonov/ridge Regularization: ($\|c\|_{B_*}^2 = c^\top B_* c$)

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

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

# Computation: Regression and Regularization

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

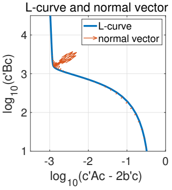
$$\mathcal{E}(\phi) = c^\top \overline{A}_n c - 2c^\top \overline{b}_n + C_N^f, \Rightarrow \widehat{\phi}_{\mathcal{H}_n} = \sum_i \widehat{c}_i\phi_i, \text{ where } \widehat{c} = \overline{A}_n^{-1}\overline{b}_n,$$

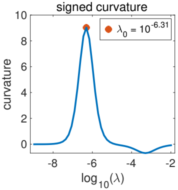
**Regularization** necessary: $\overline{A}_n$ ill-conditioned & $\overline{b}_n$: noisy or with error
Tikhonov/ridge Regularization: ($\|c\|_{B_*}^2 = c^\top B_* c$)

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

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

- Hyper-parameter $\lambda$: CV, truncated SVD, ...
  The L-curve method [Hansen00]
- Which norm $\|\cdot\|_*$?



# 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

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

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

  - $\mathcal{L}_{\overline{G}}$ is a nonnegative compact operator: $\{(\lambda_i, \psi_i)\}$, $\lambda_i \downarrow 0$

    [Open: can we make it coercive by designing data collection?]
  - $\phi^D = \mathcal{L}_{\overline{G}}\phi_{true} + \phi^{\text{error}}$

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

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

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

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

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

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

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

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

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

# Outline

**Why is DARTR good**: (1) removing error outside FSOI:

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

- DARTR: $\|\phi_{H^\perp}^{\mathrm{error}}\|_{H_G}^2 = \infty$; $\mathcal{L}_{\overline{G}}\phi_{H^\perp}^{\mathrm{error}} = 0$.

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

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

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

(2) Another metric on $H$.
What if $L^2$ is restricted to FSOI (i.e. use $I_H$)?

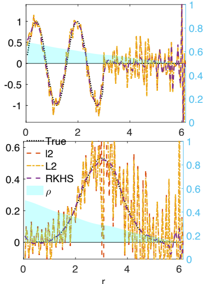
$$(\mathcal{L}_{\overline{G}} + \lambda\mathcal{L}_{\overline{G}}^{-1})^{-1}\phi^D \quad \text{v.s.} \quad (\mathcal{L}_{\overline{G}} + \lambda I_H)^{-1}\phi^D$$

Norms on $H$ for regularization: $L^2$, $H_G$, $l^2$

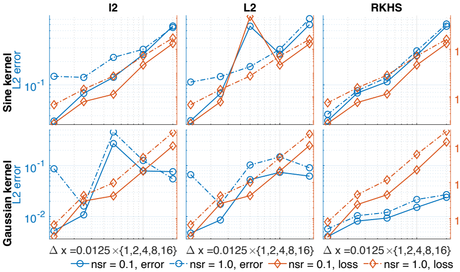
Previous numerical tests:

- DARTR has more consistent rates, but not always better.
- Depending on hyper-parameter selection

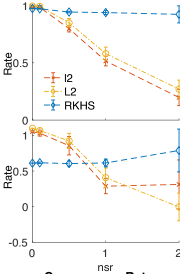
$$R_{\phi}[u] = \nabla \cdot [u(K_{\phi} * u)] = f, \quad K_{\phi} = \phi(|x|)\frac{x}{|x|}$$




Typical estimators, $\Delta x = 0.05$

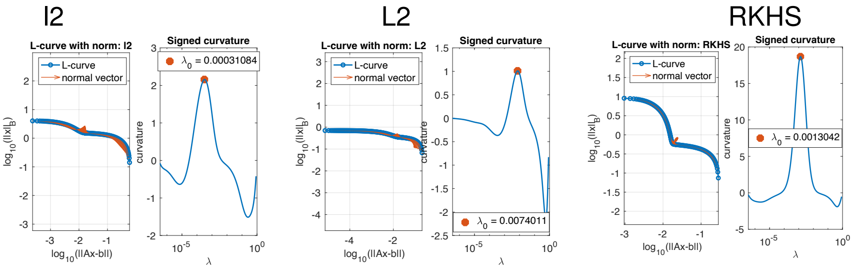



Convergence of Estimators, nsr = 0.1 & 1




Convergence Rates

# More robust L-curve



**Has DARTR been lucky in getting $\lambda_*$?**

**Can we PROVE it "better"?**

Quantitative: more accurate, robust; faster rate?

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

$$\widehat{\phi}_\lambda^{L^2_\rho} = \underset{\phi \in L^2_\rho}{\arg\min}\, \mathcal{E}(\phi) + \lambda \|\phi\|_{L^2_\rho}^2 = (\mathcal{L}_{\overline{G}} + \lambda I_H)^{-1}\phi^D,$$

$$\widehat{\phi}_\lambda^{H_G} = \underset{\phi \in H_G}{\arg\min}\, \mathcal{E}(\phi) + \lambda \|\phi\|_{H_G}^2 = (\mathcal{L}_{\overline{G}}{}^2 + \lambda I)^{-1}\mathcal{L}_{\overline{G}}\phi^D.$$

Spectral decomposition: $R_\phi[u] + \eta = f$; $\eta$ = white noise

- $\mathcal{L}_{\overline{G}}$: $\{(\lambda_k, \psi_k)\}_{k\geq 1}, \{(0, \psi_j^0)\}_{j\geq 1}$; o.n.b. of $L^2_\rho$
- $\phi^D = \mathcal{L}_{\overline{G}}\phi_* + \phi^\sigma$: $\phi^\sigma \sim \mathcal{N}(0, \sigma^2\mathcal{L}_{\overline{G}})$, $= \sum_i \sigma\xi_i\lambda_i^{1/2}\psi_i$ with $\{\xi_i\}$ iid
  [ measure on inifinite-D space: need $\mathcal{L}_{\overline{G}}$ to be of trace-class.]

$$\widehat{\phi}_\lambda^{L^2_\rho} = \underset{\phi\in L^2_\rho}{\arg\min}\, \mathcal{E}(\phi) + \lambda \|\phi\|^2_{L^2_\rho} = (\mathcal{L}_{\overline{G}} + \lambda I_H)^{-1}\phi^D,$$

$$\widehat{\phi}_\lambda^{H_G} = \underset{\phi\in H_G}{\arg\min}\, \mathcal{E}(\phi) + \lambda \|\phi\|^2_{H_G} = (\mathcal{L}_{\overline{G}}^{\ 2} + \lambda I)^{-1}\mathcal{L}_{\overline{G}}\phi^D.$$

Spectral decomposition:

- $\mathcal{L}_{\overline{G}}$: $\{(\lambda_k,\psi_k)\}_{k\geq 1}, \{(0,\psi_j^0)\}_{j\geq 1}$
- $\phi^D = \mathcal{L}_{\overline{G}}\phi_* + \phi^\sigma$: $\phi^\sigma \sim \mathcal{N}(0,\sigma^2\mathcal{L}_{\overline{G}})$, $= \sum_i \sigma\xi_i\lambda_i^{1/2}\psi_i$ with $\{\xi_i\}$ iid
- Let the true function be $\phi_* = \sum_i c_i\psi_i + \sum_j d_j\psi_j^0$

Then, the $L^2_\rho$ errors are

$$\left\|\widehat{\phi}_\lambda^{L^2_\rho} - \phi_*\right\|^2_{L^2_\rho} = \sum_i (\lambda_i+\lambda)^{-2}(\sigma\lambda_i^{1/2}\xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

$$\left\|\widehat{\phi}_\lambda^{H_G} - \phi_*\right\|^2_{L^2_\rho} = \sum_i (\lambda_i^2+\lambda)^{-2}(\sigma\lambda_i^{3/2}\xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

Which one is more accurate?

$$\left\| \widehat{\phi}_\lambda^{L^2_\rho} - \phi_* \right\|^2_{L^2_\rho} = \sum_i (\lambda_i + \lambda)^{-2} (\sigma \lambda_i^{1/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

$$\left\| \widehat{\phi}_\lambda^{H_G} - \phi_* \right\|^2_{L^2_\rho} = \sum_i (\lambda_i^2 + \lambda)^{-2} (\sigma \lambda_i^{3/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

- Too many factors: sequences $\{\lambda_i, c_i, \sigma\xi_i\}$, $\lambda$
- How to reduce the factors?

# Outline

1. Review: learning kernels
2. Why is DARTR good?
3. SNA for DARTR
4. SNA for fractional DARTR

# Small noise analysis for DARTR

$$\left\| \widehat{\phi}_\lambda^{L^2_\rho} - \phi_* \right\|_{L^2_\rho}^2 = \sum_i (\lambda_i + \lambda)^{-2} (\sigma \lambda_i^{1/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

$$\left\| \widehat{\phi}_\lambda^{H_G} - \phi_* \right\|_{L^2_\rho}^2 = \sum_i (\lambda_i^2 + \lambda)^{-2} (\sigma \lambda_i^{3/2} \xi_i - \lambda c_i)^2 + \sum_j d_j^2,$$

Assume all $d_j = 0$, i.e., $\phi_* \in$FSOI. Remove randomness by $\mathbb{E}$:

$$e^{L^2_\rho}(\lambda) = \mathbb{E} \left\| \widehat{\phi}_\lambda^{L^2_\rho} - \phi_* \right\|_{L^2_\rho}^2 = \sum_i (\lambda_i + \lambda)^{-2} (\sigma^2 \lambda_i + \lambda^2 c_i^2),$$

$$e^{H_G}(\lambda) = \mathbb{E} \left\| \widehat{\phi}_\lambda^{H_G} - \phi_* \right\|_{L^2_\rho}^2 = \sum_i (\lambda_i^2 + \lambda)^{-2} (\sigma^2 \lambda_i^3 + \lambda^2 c_i^2).$$

**Rate of convergence as $\sigma \to 0$?**

## Theorem (Small noise limit[LO23])

*Assume $\lambda_i = e^{-\theta i}$ for all $i \geq 1$ with $\theta > 0$. Let $\phi_* = \sum_i c_i \psi_i \in H$.*

(a) *When $\sup_i \lambda_i^{-1} c_i^2 < \infty$ [e.g., $\phi_* \in H_G$: $\sum_i \lambda_i^{-1} c_i^2 < \infty$] $\Rightarrow$ upper bound:*

$$\min_{\lambda>0} e^{H_G}(\lambda) \leq e^{H_G}(\sigma^2) \leq (1 + \sup_i \lambda_i^{-1} c_i^2) C_1 \sigma + O(\sigma^2),$$

*where $O(\sigma^2)$ is the big-O notation.*

(b) *Furthermore, if $\phi_*$ has $c_i^2 = \lambda_i$ ($\phi_* \in H \setminus H_G$) $\Rightarrow$ sharp rates:*

$$\lambda_* = \arg\min_{\lambda>0} e^{H_G}(\lambda) = \sigma^2, \qquad e^{H_G}(\lambda_*) = \frac{\pi}{4\theta}\sigma + O(\sigma^2);$$

$$\widetilde{\lambda}_* = \arg\min_{\lambda>0} e^{L^2_\rho}(\lambda) = \sigma + O(\sigma^2), \qquad e^{L^2_\rho}(\widetilde{\lambda}_*) = \frac{2}{\theta}\sigma + O(\sigma^2).$$

## Scheme of small noise analysis

Three steps:

Step 1: Reduce the optimization in $\lambda$ to solving an algebraic equation;

Step 2: Use integrals to approx. the series (dominating terms, small $\lambda$);

Step 3: Solve an algebraic equation for $\lambda_*$; compute optimal rate.

# Scheme of small noise analysis

Three steps:

Step 1: Reduce the optimization in $\lambda$ to solving an algebraic equation;

Step 2: Use integrals to approx. the series (dominating terms, small $\lambda$);

Step 3: Solve an algebraic equation for $\lambda_*$; compute optimal rate.

Wahba77 [Grace Wahba. Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. numerical analysis, 14(4):651–667, 1977.]

$$e(\lambda, s) := \sum_i (\lambda_i^{1+s} + \lambda)^{-2}(\sigma^2 \lambda_i^{1+2s} + \lambda^2 c_i^2)$$

$$e^{L^2_\rho}(\lambda) = \sum_i (\lambda_i + \lambda)^{-2}(\sigma^2 \lambda_i + \lambda^2 c_i^2) = e(\lambda, 0),$$

$$e^{H_G}(\lambda) = \sum_i (\lambda_i^2 + \lambda)^{-2}(\sigma^2 \lambda_i^3 + \lambda^2 c_i^2) = e(\lambda, 1).$$

Step 1: For each $s \in \{0,1\}$, $\lambda_* := \arg\min_{\lambda>0} e(\lambda, s)$ satisfies

$$\lambda = -\sigma^2 \frac{A'(\lambda; s)}{2B_1(\lambda; s)},$$

$$-\frac{1}{2}A'(\lambda; s) = \sum_i (\lambda_i^{s+1} + \lambda)^{-3}\lambda_i^{2s+1}, \quad B_1(\lambda; s) = \sum_i (\lambda_i^{s+1} + \lambda)^{-3}\lambda_i^{s+1}c_i^2.$$

Proof:

$$\begin{aligned} e(\lambda, s) : &= \sum_i (\lambda_i^{1+s} + \lambda)^{-2}(\sigma^2\lambda_i^{1+2s} + \lambda^2 c_i^2) \\ &= \sigma^2 A(\lambda, s) + \lambda^2 B(\lambda, s) \end{aligned}$$

$$0 = \frac{d}{d\lambda}e(\lambda, s) = \sigma^2 A'(\lambda, s) + 2\lambda \underbrace{\left[B(\lambda, s) + \frac{\lambda}{2}B'(\lambda, s)\right]}_{B_1(\lambda,s)}$$

$s = 1$: if $c_i^2 = \lambda_i$, then $-\frac{1}{2}A'(\lambda; s) = B_1(\lambda, s)$. Then,

$$H_G - \text{regularizer} : \lambda_* = \sigma^2, e^{H_G}(\lambda_*) = \sigma^2 A(\sigma^2, 1) + \sigma^4 B(\sigma^2, 1)$$

Step 2: estimate the dominating order of these series. $\lambda$ small, $c_i = \lambda_i$

$$-\frac{1}{2}A'(\lambda;0) = \sum_i (\lambda_i^{s+1}+\lambda)^{-3}\lambda_i^{2s+1} = \frac{(1+2\lambda)}{2\theta\lambda^2(1+\lambda)^2} + O(1),$$

$$B_1(\lambda;0) = \sum_i (\lambda_i^{s+1}+\lambda)^{-3}\lambda_i^{s+1}c_i^2 = \frac{1}{2\theta\lambda(1+\lambda)^2} + O(1)$$

$$A(\lambda;s) = \sum_i (\lambda_i^{s+1}+\lambda)^{-2}\lambda_i^{2s+1} = \frac{1}{2\theta\sqrt{\lambda}}[\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}] + O(1)$$

$$B(\lambda;s) = \sum_i (\lambda_i^{s+1}+\lambda)^{-2}c_i^2 = \frac{1}{2\theta}\lambda^{-3/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}] + O(1)$$

Basic idea:

The series = Riemann sum + O(1);    Riemann sum= $O(\lambda^{-x})$

Step 2: estimate the dominating order of these series. $\lambda$ small, $c_i = \lambda_i$

$$-\frac{1}{2}A'(\lambda;0) = \sum_i (\lambda_i^{s+1} + \lambda)^{-3}\lambda_i^{2s+1} = \frac{(1+2\lambda)}{2\theta\lambda^2(1+\lambda)^2} + O(1),$$

$$B_1(\lambda;0) = \sum_i (\lambda_i^{s+1} + \lambda)^{-3}\lambda_i^{s+1}c_i^2 = \frac{1}{2\theta\lambda(1+\lambda)^2} + O(1)$$

$$A(\lambda;s) = \sum_i (\lambda_i^{s+1} + \lambda)^{-2}\lambda_i^{2s+1} = \frac{1}{2\theta\sqrt{\lambda}}[\arctan\frac{1}{\sqrt{\lambda}} - \frac{\sqrt{\lambda}}{1+\lambda}] + O(1)$$

$$B(\lambda;s) = \sum_i (\lambda_i^{s+1} + \lambda)^{-2}c_i^2 = \frac{1}{2\theta}\lambda^{-3/2}[\arctan\frac{1}{\sqrt{\lambda}} + \frac{\sqrt{\lambda}}{1+\lambda}] + O(1)$$

Basic idea:

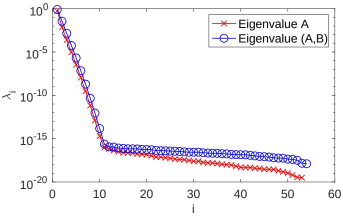
The series = Riemann sum + O(1);     Riemann sum= $O(\lambda^{-x})$

Step 3: solve the algebraic equations for $\lambda_*$ and compute $e(\lambda_*, s)$.

**Numerical tests on Fredholm equation** of the 1st kind:

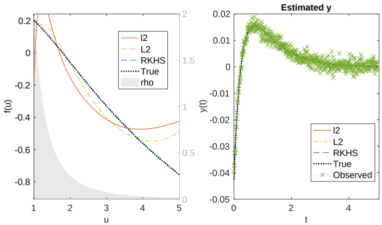
$$y(t) = \int_a^b K(s,t)\phi(s)ds + \sigma \dot{W}(t), \quad K(s,t) = s^{-2}e^{-st},\ t \in [c,d]$$

- Data: $(y(t_1), \ldots, y(t_m)) \in \mathbb{R}^m$.
- Discrete problem: $\phi$ on a mesh.
- Exponential spectrum decay

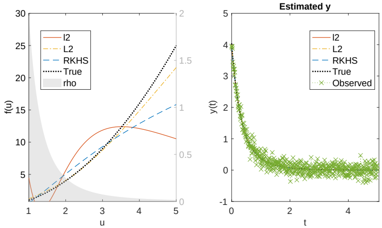



Similar to learning kernels in operators
$R_\phi[u](x) = \int \phi(|x-y|)g[u](x,y)dy$

Typical estimators when $nsr = 2$ and their recovery of the signal.




(a) True solution inside FSOI




(b) True solution outside FSOI

- All estimators recover the signal $y$ accurately (de-noising)
- $H_G$ outperforms $l^2$ and $L^2$ in (a),
  but it slightly underperforms the $L^2$ regularizer in (b).

## Convergence in small noise limit

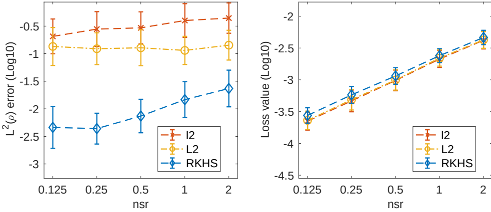
Settings:

- Mean and std in 100 realizations.
- Hyper-parameter selected from data.
- $c_i^2$ : not decaying as $\lambda_i$

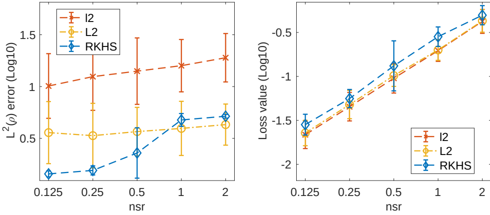
Results:

- Errors “decay" with $\sigma$
- (a): $H_G$ outperforms $l^2$ and $L^2$; (b): slightly outperforms $L^2_\rho$.

(a) $\phi_{true}$ inside FSOI




(b) $\phi_{true}$ outside FSOI



# Outline

# Fractional DARTR

## Definition (Fractional RKHS)

For $s \geq 0$, $H = \mathrm{Null}(\mathcal{L}_{\overline{G}})^{\perp}$, $\phi = \sum_{i:\lambda_i > 0} c_{i,\phi}\psi_i$:
$H_G^s = \mathcal{L}_{\overline{G}}^{s/2}(H)$ with norm $\|\phi\|_{H_G^s}^2 = \|\mathcal{L}_{\overline{G}}^{-s/2}\phi\|_{L^2_\rho}^2 = \sum_i \lambda_i^{-s} c_{i,\phi}^2$

- $s = 0$: $H_G^0 = H$; $s = 1$: $H_G^1 = H_G$
- Similar to Sobolev space when $\lambda_k = k^{-2}$?

# Fractional DARTR

## Definition (Fractional RKHS)

For $s \geq 0$, $H = \mathrm{Null}(\mathcal{L}_{\overline{G}})^{\perp}$, $\phi = \sum_{i:\lambda_i>0} c_{i,\phi}\psi_i$:

$H_G^s = \mathcal{L}_{\overline{G}}^{s/2}(H)$ with norm $\|\phi\|_{H_G^s}^2 = \|\mathcal{L}_{\overline{G}}^{-s/2}\phi\|_{L_\rho^2}^2 = \sum_i \lambda_i^{-s} c_{i,\phi}^2$

- $s = 0$: $H_G^0 = H$; $s = 1$: $H_G^1 = H_G$
- Similar to Sobolev space when $\lambda_k = k^{-2}$?

**Fractional DARTR:**

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

- $s$ control the smoothness, $\lambda$ control regu. strength
- Should the best $s$ be the regularity of $\phi_{true} \in H_G^r$?

## Theorem (Rates in small noise limit[LL23])

- *Spectrum decay:* $\lambda_k = p_i f(i)$ *with* $p_i \in [a,b] \subset \mathbb{R}^+$ *and* $f(x) = x^{-\theta}$ *or* $e^{-\theta(x-1)}$ *(denote* $\beta = \theta^{-1} + 1$ *or* $1$*)*
- $r$*-smoothness of* $\phi_{true}$*:* $\phi_* = \sum_i c_i \psi_i \in L^2_\rho$ *with* $|c_i| = \lambda_i^r$, $r > 0$.

*Then, the minimal* $H^s_G$*-regularizer's error satisfies*

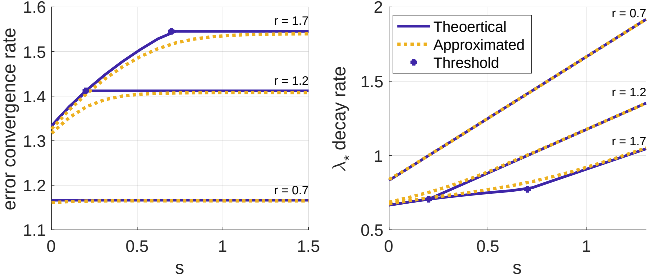
$$\lambda_* \simeq \begin{cases} \sigma^{\frac{2s+2}{2r+1}}, & s > r - \frac{\beta+1}{2}, \\ \sigma^{\frac{2s+2}{2s+2+\beta}}, & s < r - \frac{\beta+1}{2}; \end{cases} \quad e(\lambda_*; s) \simeq \begin{cases} \sigma^{2-\frac{2\beta}{2r+1}}, & s > r - \frac{\beta+1}{2}, \\ \sigma^{2-\frac{2\beta}{2s+2+\beta}}, & s < r - \frac{\beta+1}{2}. \end{cases}$$

- Optimal rate depends on all factors $(s, \theta, r)$!
- Proof using the SNA-scheme
- Not optimal near the threshold $s = r - \frac{\beta+1}{2}$ (no algebraic equation due to a log-term)

Should $s = r$ when $H^s_G$-regularizer and $\phi_{true} \in H^r_G$?

## Should $s = r$ when $H_G^s$-regularizer and $\phi_{true} \in H_G^r$?

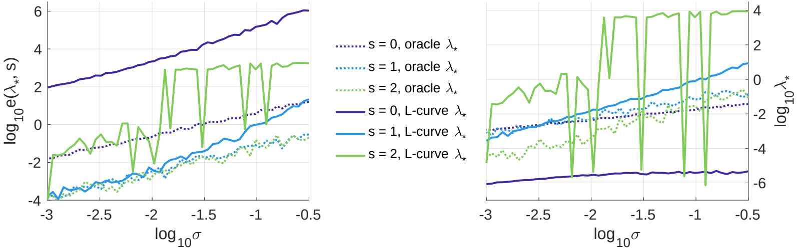
Settings: $f(x) = e^{-1.5x}$, $\beta = 1$




- Over-smoothing OK (according to the rate)
- Trouble in the selection of $\lambda_*$ when $s$ large

## Over-smoothing makes it difficult to select the optimal $\lambda_*$

Settings: $f(x) \approx x^{-4}$, $\beta = \frac{5}{4}$, $r = 1.5$.




- Over-smoothing ($s = 2$):
  difficult to select $\lambda_*$(right), leading to a relatively large error (left).
- Under-smoothing ($s = 0$):
  optimal $\lambda_*$ too small (right); leading to large error (left)
- Properly regularization with $s = 1$:
  $\lambda_*$ close to the oracle ones, leading accurate estimators

# Summary

Compare regularization norms: small noise analysis

- Practice: too many factors to analyze
- Small noise analysis:
  - Reduce the complexity to rate in $\sigma \to 0$
    - spectrum decay
    - smoothness: fractional space $\mathcal{L}_{\overline{G}}^{s}$
    - Oracle $\lambda_*$ minimizing $L^2$-error
  - A simple scheme: Riemann sum + algebraic equations

A surprising insight:
Over-smoothing OK in theory; trouble in optimal $\lambda_*$ selection

## Future directions

Inverse problems $\leftrightarrow$ Learning with nonlocal dependence

- Convergence: $\Delta x, N$? Minimax rate?
- Jointly select $(s, \lambda)$ in computation? Iterative DARTR?
- Automatic kernel for Gaussian Process/Kernel Regression?

## Future directions

Inverse problems $\leftrightarrow$ Learning with nonlocal dependence

- Convergence: $\Delta x, N$? Minimax rate?
- Jointly select $(s, \lambda)$ in computation? Iterative DARTR?
- Automatic kernel for Gaussian Process/Kernel Regression?