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

Lecture 5. DARTR: data adaptive RKHS Tikhonov regularization

Learning interaction kernel $K_\phi(x-y) = \phi(|x-y|)\frac{x-y}{|x-y|}$

$$dX_t^i = \frac{1}{N}\sum_{j=1}^{N} K_\phi(X_t^j - X_t^i)dt + \sqrt{2\nu}dB_t^i \quad \Leftrightarrow R_\phi(\mathbf{X}_t) = \dot{\mathbf{X}}_t - \sqrt{2\nu}\dot{\mathbf{B}}_t$$

$$\partial_t u = \nu\Delta u + \nabla\cdot[u(K_\phi * u)] \quad \Leftrightarrow R_\phi[u(\cdot,t)] = f(\cdot,t)$$

- N small: M trajectories $\{\mathbf{X}^{(m)}_{t_1:t_L}\}_{m=1}^M \Rightarrow$ statistical learning
- $N=\infty$: $\{u_N(x,t_l) = \frac{1}{N}\sum_i \delta_{X^i_{t_l}}(x)\}$ or $\{u(x_m,t_l)\}_{m,l} \Rightarrow$ inverse problem

Nonparametric regression: $\phi = \sum_{k=1}^n c_k\phi_k$

$$A_n c = b_n$$

Ill-conditioned A_n $\Rightarrow$ **Regularization**

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

How to choose $\|c\|_*^2$?

1. Learning kernels
2. Regression and regularization
3. Identifiability and DARTR
4. Adaptive prior for Bayesian inverse

- LLA22: Lu, Lang, and An. MSML22.
- LAY22: Lu, An and Yu. arXiv2205
- CLLW22: Chada, Lang, L.& Wang.arXiv2212

Outline

1. Learning kernels
2. Regression and regularization
3. Identifiability and DARTR
4. Adaptive prior for Bayesian inverse

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

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

$$R_\phi[\mathbf{X}_t] = \Big[-\frac{1}{N}\sum_{j=1}^N K_\phi(X_t^i - X_t^j)\Big]_i = \dot{\mathbf{X}}_t + \dot{\mathbf{W}}_t, \qquad u = \frac{1}{N}\sum_{i=1}^N \delta(X_t^i - x)$$

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

Outline

1. Learning kernels
2. Regression and regularization
3. Identifiability and DARTR
4. Adaptive prior for Bayesian inverse

Nonparametric regression

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

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

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

- $\overline{A}_n^{-1}$: 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
- $\overline{b}_n$: noisy or with numerical error

Tikhonov/ridge Regularization: $(\|c\|_{B_*}^2 = c^\top B_* c)$

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

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

Regularization

Regularization is necessary:

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

Tikhonov/ridge Regularization: ($\|c\|^2_{B_*} = c^\top B_* c$)

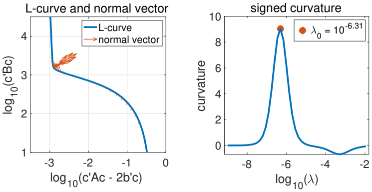
$$\mathcal{E}_\lambda(\phi) = \mathcal{E}(\phi) + \lambda\|\phi\|^2_* \Rightarrow c^\top \overline{A}_n c - 2\overline{b}_n^\top c + \lambda\|c\|^2_{B_*}$$

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

- hyper-parameter λ: CV, truncated SVD, ...
- The L-curve method [Hansen00]

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

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



Regularization

Regularization is necessary:

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

Tikhonov/ridge Regularization: ($\|c\|^2_{B_*} = c^\top B_* c$)

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

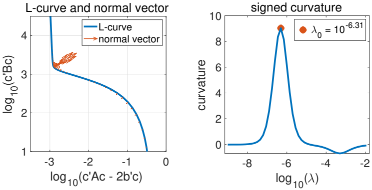
$$\widehat{\phi}^\lambda_{\mathcal{H}_n} = \sum_i \widehat{c}^\lambda_i \phi_i, \qquad \text{where } \widehat{c} = (\overline{A}_n + \lambda B_*)^{-1}\overline{b}_n,$$

- hyper-parameter λ: CV, truncated SVD, ...
- The L-curve method [Hansen00]

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

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

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



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?

Outline

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

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

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

  - $\mathcal{L}_{\overline{G}}$ 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: $\rho(dr) \qquad \Rightarrow \phi \in L^2(\rho)$

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

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

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

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

  - $\mathcal{L}_{\overline{G}}$ 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**):

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

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

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

A new task for Regularization:

ensure that the learning takes place in the FSOI

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

- $\overline{G}$ ⇒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: $\|\phi\|^2_{H_G}$

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

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 $C = \sum_i \phi_i \otimes \phi_i$ or $C = I$

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

DARTR: computation

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): $A_n V = B V \Lambda$
- $B_{rkhs} = (V \Lambda V^\top)^{-1}$: avoid inverse $\downarrow$
  Set $D = B_n^{-1} A_n^{1/2}$, then, $\widehat{c}_\lambda = D(D^\top A_n D + \lambda I)^{-1} D^\top b_n$
- L-curve to select λ_*

Connecting computation with theory

$$\widehat{c}_\lambda = (A_n + \lambda_* B_{rkhs})^{-1} b_n \quad \leftrightarrow \qquad \widehat{\phi} = \mathcal{L}_{\overline{G}}{}^{-1} \phi^D$$

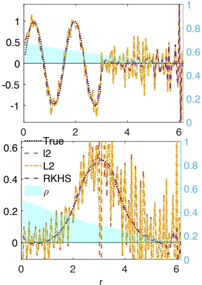
Theorem: If $\mathcal{L}_{\overline{G}}(L^2(\rho)) \subset \mathcal{H}$, then, $\mathcal{L}_{\overline{G}}$ eigenvalues $\leftrightarrow$ GEP (A_n, B_n).

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

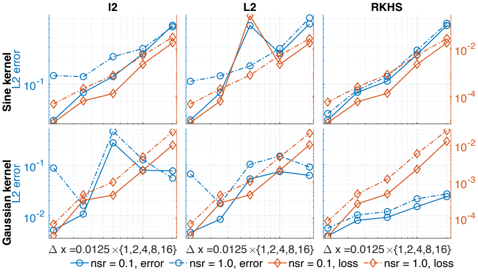
Interaction kernel in a nonlinear operator

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

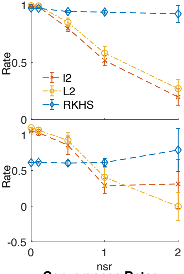
- Recover kernel from discrete noisy data
- Robust in accuracy, consistent rates as mesh refines



Typical estimators, $\Delta x = 0.05$



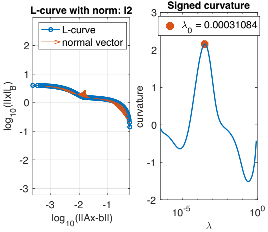
Convergence of Estimators, nsr = 0.1 & 1



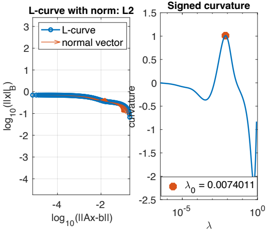
Convergence Rates

More robust L-curve

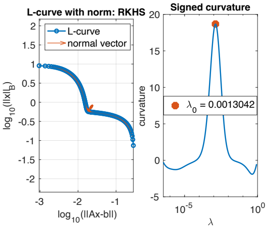
l2



L2

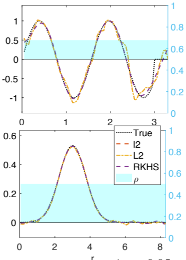


RKHS

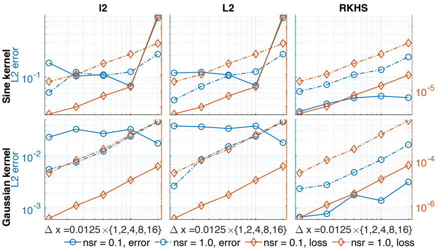


Linear integral operator:

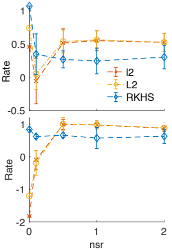
$$R_\phi[u](x) = \int_\Omega \phi(|y-x|)u(y)dy = f(x).$$



Typical estimators, $\Delta x = 0.05$



Convergence of Estimators, nsr = 0.1 & 1



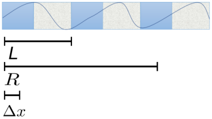
Convergence Rates

Robust in accuracy, consistent rates

Homogenization of wave propagation in meta-material

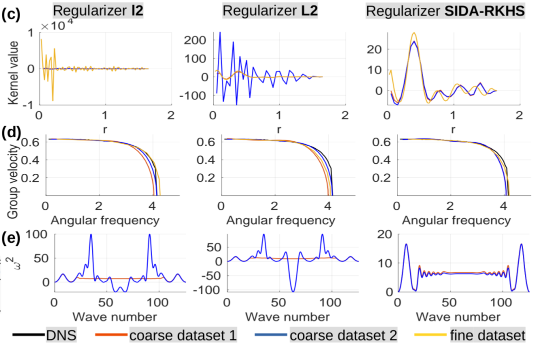
- heterogeneous bar with microstructure + DNS $\Rightarrow$ Data
- Homogenization: $R_\phi[u] = \int_\Omega \phi(|y|)[u(x+y) - u(x)]dy = \partial_{tt}u - g.$

(a) Wave propagation in a heterogeneous bar



(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

Outline

Bayesian inverse problem

Variational inverse problem: ill-posed

$$\widehat{\phi} = \arg\min_{\phi} \mathcal{E}(\phi) = \langle \mathcal{L}_{\overline{G}}\phi, \phi\rangle_{L^2(\rho)} - 2\langle \phi^D, \phi\rangle_{L^2(\rho)} + C.$$

Finite dimensional case:

- **Prior** $\mathcal{N}(0, \mathcal{Q}_0)$ with $\mathcal{Q}_0 > 0$: $\frac{d\pi_0(\phi)}{d\phi} \propto e^{-\frac{1}{2}\langle \phi, \mathcal{Q}_0^{-1}\phi\rangle_{L^2(\rho)}}$
- **Likelihood** of data: $\pi_L \sim \mathcal{N}(\mathcal{L}_{\overline{G}}{}^{-1}\phi^D, \sigma_\eta^2 \mathcal{L}_{\overline{G}}{}^{-1})$

$$\frac{d\pi_L(\phi)}{d\phi} \propto \exp(-\frac{1}{2\sigma_\eta^2}\mathcal{E}(\phi))$$

- **Posterior** = Gaussian $\mathcal{N}(\mu_1, \mathcal{Q}_1)$ well-posed

$$\frac{d\pi_1(\phi)}{d\phi} \propto \exp\left(-\frac{1}{2}[\sigma_\eta^{-2}\mathcal{E}(\phi) + \langle \mathcal{Q}_0^{-1}\phi, \phi\rangle_{L^2(\rho)}]\right)$$

$\mu_1 = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}\phi^D$, and $\mathcal{Q}_1 = \sigma_\eta^2(\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}$.

Infinite-dimensional prior and posterior

$\phi \in L^2_\rho$, infinite-dimensional:

- **Prior** $\mathcal{N}(0, \mathcal{Q}_0)$, where $\mathcal{Q}_0 > 0$, trace-class operator;
- **Posterior** $\mathcal{N}(\mu_1, \mathcal{Q}_1)$: Radon–Nikodym derivative w.r.t the prior

$$\frac{d\pi_1}{d\pi_0} \propto \exp(-\frac{1}{2}\sigma_\eta^{-2}\mathcal{E}(\phi)) = \exp\left(-\frac{1}{2}\sigma_\eta^{-2}[\langle \mathcal{L}_{\overline{G}}\phi, \phi\rangle_{L^2(\rho)} - 2\langle \phi^D, \phi\rangle_{L^2(\rho)} + C]\right),$$

$$\mu_1 = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}\phi^D, \text{ and } \mathcal{Q}_1 = \sigma_\eta^2(\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}.$$

Is it a good idea to choose $\mathcal{Q}_0 > 0$?

Risk in a non-degenerate prior

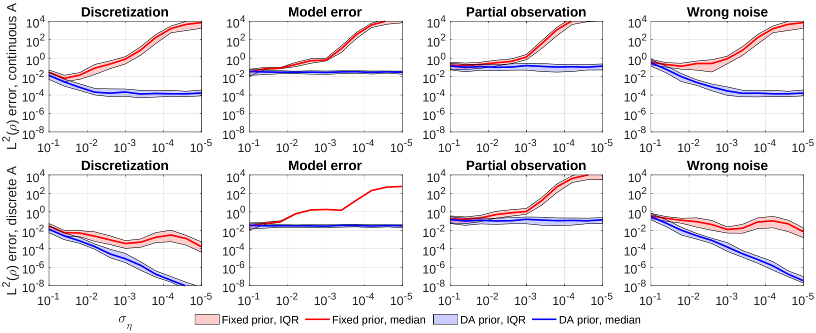
Theorem ([CLLW22])

A non-degenerate prior has the risk of leading to a divergent posterior mean in the small noise limit.
I.e., $\lim_{\sigma_\eta \to 0}(\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}\phi^{D}$ "$=$" ∞ *under Assumption (next page).*

Risk in a non-degenerate prior

Theorem ([CLLW22])

A non-degenerate prior has the risk of leading to a divergent posterior mean in the small noise limit.
I.e., $\lim_{\sigma_\eta\to 0}(\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1}\phi^D \text{``}=\text{''}\infty$ *under Assumption (next page).*



Assumptions: $\mathcal{L}_{\overline{G}} \sim \{(\lambda_k, \psi_k)\}$

- *Finite rank* $\mathcal{L}_{\overline{G}}$:
  $\lambda_k > 0$ for $1 \leq k \leq K$, $\lambda_k = 0$ for $k > K$. $\dim(L^2(\rho)) > K$.
  As a result, the FSOI $H = \mathrm{span}\{\psi_i\}_{i=1}^{K} \subsetneq L^2_\rho$.
- *Prior covariance commutes with* $\mathcal{L}_{\overline{G}}$:
  The prior $\mathcal{N}(0, \mathcal{Q}_0)$ satisfies $\mathcal{Q}_0 \psi_i = r_i \psi_i$ with $r_i > 0$ for all i.
- *Error outside of FSOI.*

$$\phi^{D} = \mathcal{L}_{\overline{G}} \phi_{true} + \epsilon^{\xi} + \epsilon^{\eta},$$

The model error $\epsilon^{\xi} = \sum_i \epsilon_i^{\xi} \psi_i$ has $\epsilon_{i_0}^{\xi} \neq 0$ with $i_0 > K$.

Note: $\epsilon^{\eta} \sim \mathcal{N}(0, \sigma_\eta^2 \mathcal{L}_{\overline{G}})$ comes from noise.

Proof:

$$\phi^D = \sum_i \phi_i^D \psi_i, \text{ with } \phi_i^D = \lambda_i \phi_{true,i} + \sigma_\eta \lambda_i^{1/2} \epsilon_i^\eta + \epsilon_i^\xi.$$

The posterior mean $\mu_1 = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1} \phi^D$ becomes

$$\mu_1 = \sum_i \left(\lambda_i + \sigma_\eta^2 r_i^{-1}\right)^{-1} \phi_i^D \psi_i = \sum_{i=1}^K \left(\lambda_i + \sigma_\eta^2 r_i^{-1}\right)^{-1} \phi_i^D \psi_i + \sum_{i>K} \sigma_\eta^{-2} r_i \epsilon_i^\xi \psi_i.$$

The first term $\to \sum_{1\le i\le K} \left(\phi_{true,i} + \lambda_i^{-1} \epsilon_i^\xi\right) \psi_i$

But the 2nd term $\sum_{i>K} \sigma_\eta^{-2} r_i \epsilon_i^\xi \psi_i$ diverges because $\exists \epsilon_{i_0}^\xi \neq 0$.

Data-adaptive prior

Fixed prior $\phi \sim \pi_0 = \mathcal{N}(0, \mathcal{Q}_0)$, $\pi_1 = \mathcal{N}(\mu_1, \mathcal{Q}_1)$

$$\mu_1 = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1} \phi^{\scriptscriptstyle D}$$

Data-adaptive prior $\phi \sim \pi_0 = \mathcal{N}(0, \lambda_*^{-1} \mathcal{L}_{\overline{G}})$, $\pi_1 = \mathcal{N}(\mu_1, \mathcal{Q}_1^{\scriptscriptstyle D})$

$$\mu_1^{\scriptscriptstyle D} = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}^{\;-1})^{-1} \phi^{\scriptscriptstyle D}$$

$$\mathcal{Q}_1^{\scriptscriptstyle D} = \sigma_\eta^2 (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}^{\;-1})^{-1}$$

Data-adaptive prior

Fixed prior $\phi \sim \pi_0 = \mathcal{N}(0, \mathcal{Q}_0)$, $\pi_1 = \mathcal{N}(\mu_1, \mathcal{Q}_1)$

$$\mu_1 = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \mathcal{Q}_0^{-1})^{-1} \phi^{D}$$

Data-adaptive prior $\phi \sim \pi_0 = \mathcal{N}(0, \lambda_*^{-1} \mathcal{L}_{\overline{G}})$, $\pi_1 = \mathcal{N}(\mu_1, \mathcal{Q}_1^{D})$

$$\mu_1^{D} = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}^{-1})^{-1} \phi^{D}$$

$$\mathcal{Q}_1^{D} = \sigma_\eta^2 (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}^{-1})^{-1}$$

Theorem (Small noise limit of the posterior mean [CLLW22])

Fixed prior with $\mathcal{Q}_0 = I_d$: blows up;
Data-adaptive prior: stable and convergent

Proof With $\sigma_\eta^2 \lambda_* > 0$:

$$(\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}{}^{-1})^{-1} \psi_i = (\mathcal{L}_{\overline{G}}{}^2 + \sigma_\eta^2 \lambda_* I)^{-1} \mathcal{L}_{\overline{G}} \psi_i$$
$$= \frac{\lambda_i}{\lambda_i^2 + \sigma_\eta^2 \lambda_*} \psi_i = 0\,, \text{ if } i \geq K+1\,.$$

Then, the posterior mean is:

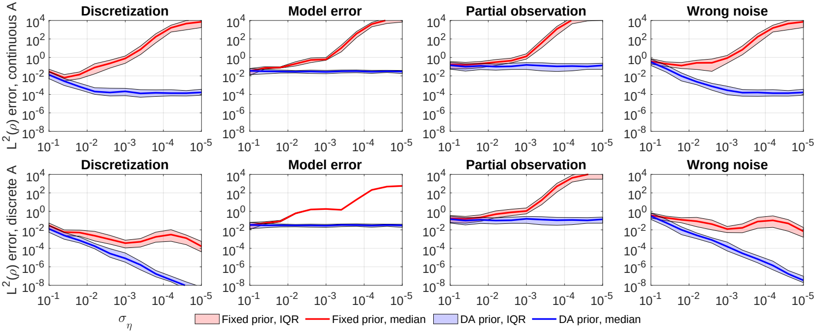
$$\mu_1^D = (\mathcal{L}_{\overline{G}} + \sigma_\eta^2 \lambda_* \mathcal{L}_{\overline{G}}{}^{-1})^{-1} \phi^D = \sum_{1 \leq i \leq K} \left(\lambda_i + \sigma_\eta^2 \lambda_* \lambda_i^{-1} \right)^{-1} \phi_i^D \psi_i.$$

The limit exist as $\sigma_\eta \to 0$:
(recall that $\phi^D = \sum_i \phi_i^D \psi_i$, with $\phi_i^D = \lambda_i \phi_{true,i} + \sigma_\eta \lambda_i^{1/2} \epsilon_i^\eta + \epsilon_i^\xi$)

$$\lim_{\sigma_\eta \to 0} \mu_1^D = \sum_{i=1}^{K} \left(\phi_{true,i} + \lambda_i^{-1} \epsilon_i^\xi \right) \psi_i.$$

Small noise limit of posterior mean

Integral operator, ϕ_{true} inside the FSOI:

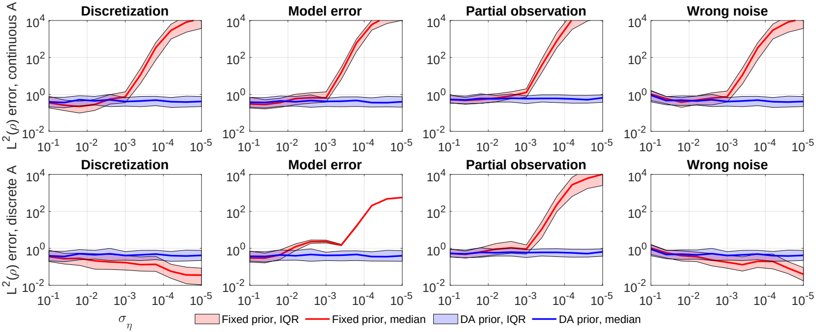


Fixed prior: blowup

Data-adaptive prior: stable and convergent

Small noise limit of posterior mean

Integral operator, ϕ_{true} outside the FSOI:



Fixed prior: blowup

Data-adaptive prior: stable and convergent

Summary

Learning kernels in operators:

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

Nonlocal dependence

- Identifiability: data-dependent FSOI
- DARTR: data adaptive RKHR Tikhonov-Reg
  - Synthetic data: convergent, robust to noise
  - Homogenization: resolution-independent
- Small noise analysis in Bayesian

Next:

How do we theoretically compare different regularizations?

Small noise analysis[LO23,LL23]

Future directions

Inverse problems $\leftrightarrow$ Learning with nonlocal dependence

- Iterative methods for DARTR
- Convergence: $\Delta x, N$
- Regularization for ML: $\|\phi_\theta\|^2_{rkhs}$, not $\|\theta\|$

Next:

How do we theoretically compare different regularizations?

Small noise analysis[LO23,LL23]

Future directions

Inverse problems $\leftrightarrow$ Learning with nonlocal dependence

- Iterative methods for DARTR
- Convergence: $\Delta x, N$
- Regularization for ML: $\|\phi_\theta\|^2_{rkhs}$, not $\|\theta\|$

References

(@ http://www.math.jhu.edu/~feilu)

- LLA22: Lu, Lang, and An. MSML22. (Matlab code available)
- LAY22: Lu, An and Yu. arXiv2205
- CLLW22: Chada, Lang, Lu, and Wang. arXiv2212