I: Nonparametric learning of kernels in operators

Fei Lu

Department of Mathematics, Johns Hopkins University
feilu@math.jhu.edu

Fall 2023

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

1. Overview and a review of classical learning theory

1. An overview with examples

2. Nonparametric regression and main results

3. Classical learning theory

4. Applying classical learning theory to IPS

Outline

1. An overview with examples

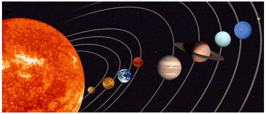
2. Nonparametric regression and main results

3. Classical learning theory

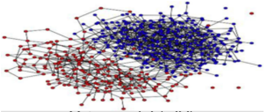
4. Applying classical learning theory to IPS

A motivating example

What is the law of interaction ?



Popkin. Nature(2016)



Voter model (wiki)

$$\ddot{X}_t^i = \frac{1}{N}\sum_{j=1, j\neq i}^{N} m_j K_\phi(X_t^j - X_t^i),$$

$$K_\phi(x-y) = \nabla_x[\Phi(|x-y|)] = \phi(|x-y|)\frac{x-y}{|x-y|}.$$

- Newton's law of gravity $\phi(r) = \frac{c_1}{r^2}$
- Lennard-Jones potential: $\Phi(r) = \frac{c_1}{r^{12}} - \frac{c_2}{r^6}$.

- flocking birds, schooling fish, migrating cells, ...?
- opinions, people, agents in social network, ...? [a]

Infer the interaction kernel from data?

[a] (1) Cucker+Smale: On the mathematics of emergence. 2007. (2) Vicsek+Zafeiris: Collective motion. 2012. (3) Motsch+Tadmor: Heterophilious Dynamics Enhances Consensus. 2014 ...

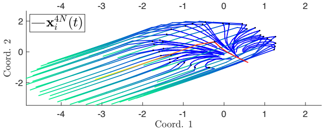
Learning the interaction kernel ϕ

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

$$K_\phi(x,y) = \phi(|x-y|)\frac{x-y}{|x-y|}$$

Finite N:

- Data: M trajectories of particles $\{\mathbf{X}_{t_1:t_L}^{(m)}\}_{m=1}^{M}$
- Statistical learning

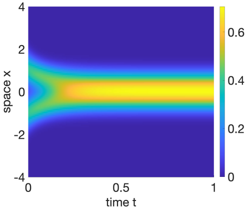


Large N ($>> 1$)

- Data: density of particles
  $\{u_N(x,t_l) = N^{-1}\sum_i \delta(X_{t_l}^i - x)\}$ or $\{u(x_m,t_l)\}_{m,l}$

$$\partial_t u = \nu\Delta u + \nabla\cdot[u(K_\phi * u)]$$

- Inverse problem for a PDE



Learning kernels in operators:

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

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

$$\boxed{\text{Infer } \phi \text{ in} \quad R_\phi[u] = f \quad \text{from data } \mathcal{D} = \{(u_k, f_k)\}_{k=1}^{M}}$$

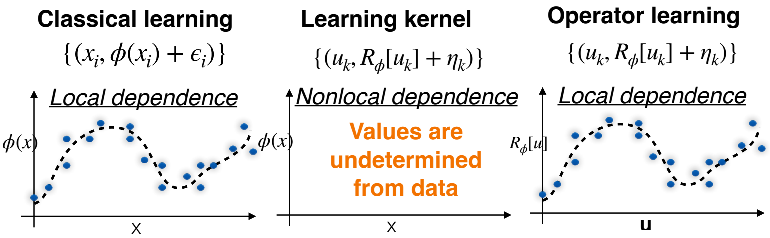
- R_ϕ linear/nonlinear in u, but linear in ϕ
- Other examples: .
  - Integral/nonlocal operators,...
  - Memory kernel in GLE,...
  - Unsupervised regression

What is new from

- classical learning $\{(x_i, y_i)\}_{i=1}^{M} \Rightarrow y = \phi(x)$?
- operator learning $\{(u_k, f_k)\}_{k=1}^{M} \Rightarrow f = R[u]$?

What is new from

- classical learning $\{(x_i, y_i)\}_{i=1}^M \Rightarrow y = \phi(x)$?
- operator learning $\{(u_k, f_k)\}_{k=1}^M \Rightarrow f = R[u]$?



Outline

1. An overview with examples

2. Nonparametric regression and main results

3. Classical learning theory

4. Applying classical learning theory to IPS

Nonparametric regression in computation

$$\dot{X}_t^i = -\frac{1}{N}\sum_{j=1}^{N} \phi(|X_t^i - X_t^j|)\frac{X_t^i - X_t^j}{|X_t^i - X_t^j|},\ i = 1,\ldots,N \quad \Leftrightarrow \quad \dot{\mathbf{X}}_t = R_{\phi}(\mathbf{X}_t)$$

Given: data $\{\mathbf{X}_{[0,T]}^{(m)}\}_{m=1}^{M}$ from ϕ_{true}. Goal: estimate ϕ.

Nonparametric regression in computation

$$\dot{X}_t^i = -\frac{1}{N}\sum_{j=1}^{N} \phi(|X_t^i - X_t^j|)\frac{X_t^i - X_t^j}{|X_t^i - X_t^j|},\ i = 1,\dots,N \quad \Leftrightarrow \quad \dot{\mathbf{X}}_t = R_\phi(\mathbf{X}_t)$$

Given: data $\{\mathbf{X}_{[0,T]}^{(m)}\}_{m=1}^{M}$ from ϕ_{true}. Goal: estimate ϕ.

Variational approach: $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$

$$\widehat{\phi}_{n,M} = \underset{\phi\in\mathcal{H}_n}{\arg\min}\, \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt$$

Linearity in ϕ: $R_{\alpha\phi+\beta\psi}(\mathbf{X}) = \alpha R_\phi(\mathbf{X}) + \beta R_\psi(\mathbf{X})$

$$\phi = \sum_{i=1}^{n} c_i\phi_i, \quad \mathcal{E}_M(\phi) = \mathcal{E}_M(c) = c^\top A_{n,M}c - 2c^\top b_{n,M} + Const.$$

$$\nabla\mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1}b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i\phi_i$$

Fundamental Issues

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

$$\widehat{\phi}_{n,M} = \underset{\phi \in \mathcal{H}_n}{\arg\min}\, \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_{\phi}(\mathbf{X}_t^{(m)})\right|^2 dt = c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.$$

$$\nabla \mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i \phi_i$$

Fundamental Issues

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

$$\widehat{\phi}_{n,M} = \underset{\phi\in\mathcal{H}_n}{\arg\min}\ \mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt = c^\top A_{n,M}c - 2c^\top b_{n,M} + Const.$$

$$\nabla\mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1}b_{n,M} \Rightarrow \quad \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i\phi_i$$

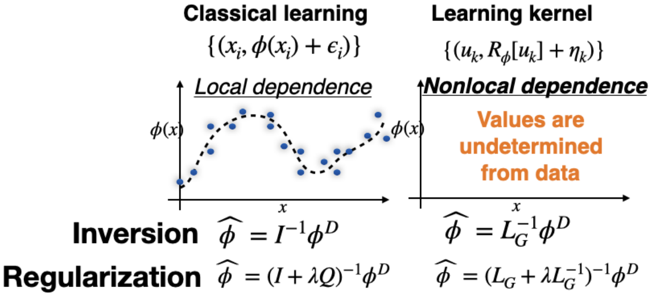
- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1}b_{n,M}$ stable?
- Identifiability of ϕ_{true}?
- Convergence of $\phi_{n,M}$? Minimax rate $\mathbb{E}\|\phi_{n_M} - \phi_{true}\|^2 \sim M^{-\frac{2s}{2s+1}}$?

Nonparametric regression/learning $\downarrow$

Main results

Large sample limit:

$$\mathcal{E}_M(\phi) \xrightarrow{M\to\infty} \mathcal{E}_\infty(\phi) = \langle L_G\phi, \phi\rangle - 2\langle \phi^D, \phi\rangle + Const$$



Main results

$$\mathcal{E}_M(\phi) \xrightarrow{M\to\infty} \mathcal{E}_\infty(\phi) = \langle L_G\phi, \phi\rangle - 2\langle \phi^D, \phi\rangle + Const$$

- With coercivity condition ($L_G \geq c_{\mathcal{H}} I$), $N < \infty$ particles:
  - Well-posed, identifiable, Minimax rate: $M^{-2s/(2s+1)}$
  - deterministic/stochastic systems, homo-/hetero-geneous systems: [LZTM19pnas, LMT19jmlr,LMT21foc];
  - Coercivity condition: partial results in [LLMTZ21spa,LL20]
- Without coercivity condition (L_G compact): $N = \infty$
  - Ill-posed/ill-defined: regularization necessary (open in computation)
  - Minimax rate: depends on the spectrum of L_G (open)
  - Construction of loss function — mean-field equation [LangLu21]

Outline

1. An overview with examples

2. Nonparametric regression and main results

3. Classical learning theory

4. Applying classical learning theory to IPS

Classical learning theory: a brief review

A brief review of relevant elements.

- Cucker-Smale2001: On the Mathematical Foundations of Learning.
- László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. A distribution-free theory of nonparametric regression. Springer Science & Business Media, 2006.
- AB Tsybakov. Introduction to nonparametric estimation. Springer 2008.

Given: Data$\{(X_m, Y_m)\}_{m=1}^{M} \sim (X, Y), \mathbb{R}^1$ random variables.

Goal: find f s.t. $Y = f(X)$ “best fit" the data.

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] \approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^{M} |Y_m - f(X_m)|^2$$

Given: Data$\{(X_m, Y_m)\}_{m=1}^{M} \sim (X, Y), \mathbb{R}^1$ random variables.

Goal: find f s.t. $Y = f(X)$ “best fit" the data.

$$\mathcal{E}(f) = \mathbb{E}[|Y - f(X)|^2] \approx \mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^{M} |Y_m - f(X_m)|^2$$

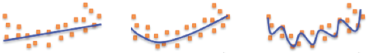
- Function space: $L^2(\rho_X)$. Best fit $f_*(x) = \mathbb{E}[Y|X = x] = \underset{f\in L^2(\rho)}{\arg\min}\, \mathcal{E}(f)$.
- Identifiability: if $Y = f_{true}(X) + \xi$ with ξ mean zero square integrable , then $f_* = f_{true}$ in $L^2(\rho_X)$.

Nonparametric Regression:

$\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n, f = \sum_{i=1}^n c_i\phi_i,$

$$\nabla \mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \quad \widehat{f}_{n,M} = \sum_i \widehat{c}_i \phi_i.$$

- $A_{n,M} \approx \mathbb{E}[\phi_i(X)\phi_j(X)] \Rightarrow$ Choose $\{\phi_i\}$ ONB in $L^2(\rho_X)$.
- $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$ with $n = n_M$ TBD



Underfitting Balanced Overfitting

Examples of hypothesis spaces
  - Finite-D with basis: (trig-)polynomials, B-splines, wavelets, ...
  - RKHS: $\phi_i = K(x_i, \cdot)$ with preselected K and $\{x_i\}_{i=1}^n$
  - May consider only a bounded set.
- Convergence of $\widehat{f}_{n_M,M}$?

- **Non-asymptotic: probabilistic bound**

  How many samples do we need to assert, with a confidence greater than $1-\delta$, that $\|\widehat{f}_{\mathcal{H}_n,M} - f_*\|_2^2 \leq \epsilon$?

  i.e., find $M_{\delta,\epsilon}$ such that $\forall M \geq M_{\delta,\epsilon}, \quad \mathbb{P}\left(\|\widehat{f}_{\mathcal{H}_n,M} - f_*\|_2^2 \geq \epsilon\right) \leq \delta$.

- **Asymptotic: Minimax rate of convergence as $M \to \infty$**

$$\mathbb{E}\|\widehat{f}_{n_M} - f_*\|_2^2 \sim M^{-\frac{2s}{2s+1}},$$

  with s being the Holder-continuity exponent of f_*.

Non-asymptotic: probabilistic bound

Find $M_{\delta,\epsilon}$ such that $\forall M \geq M_{\delta,\epsilon}$, $\quad \mathbb{P}\left(\|\widehat{f}_{n_M} - f_*\|_2^2 \leq \epsilon \right) > 1 - \delta$.

Probabilistic bounds – Concentration inequalities

Let $\{\xi_i\}_{i=1}^M$ be iid samples of ξ, a r.v. with mean μ and variance σ.

- Bernstein: if $|\xi - \mu| \leq K$ a.s., then $\forall \epsilon > 0$,
$$\mathbb{P}\left(\left| \frac{1}{M} \sum_{i=1}^M \xi_i - \mu \right| \geq \epsilon \right) \leq 2 \exp\left(-\frac{M\epsilon^2}{2\sigma^2 + \frac{2}{3}K\epsilon}\right)$$
- Hoeffding:
$$\mathbb{P}\left(\left| \frac{1}{M} \sum_{i=1}^M \xi_i - \mu \right| \geq \epsilon \right) \leq 2 \exp\left(-\frac{M\epsilon^2}{2K^2}\right)$$

Non-asymptotic: probabilistic bound

Find $M_{\delta,\epsilon}$ such that $\forall M \geq M_{\delta,\epsilon}$, $\quad \mathbb{P}\left(\|\widehat{f}_{n_M} - f_*\|_2^2 \leq \epsilon \right) > 1 - \delta$.

Probabilistic bounds – Concentration inequalities

Let $\{\xi_i\}_{i=1}^M$ be iid samples of ξ, a r.v. with mean μ and variance σ.

- Bernstein: if $|\xi - \mu| \leq K$ a.s., then $\forall \epsilon > 0$,
$$\mathbb{P}\left(\left| \frac{1}{M} \sum_{i=1}^M \xi_i - \mu \right| \geq \epsilon \right) \leq 2\exp\left(-\frac{M\epsilon^2}{2\sigma^2 + \frac{2}{3}K\epsilon}\right)$$
- Hoeffding:
$$\mathbb{P}\left(\left| \frac{1}{M} \sum_{i=1}^M \xi_i - \mu \right| \geq \epsilon \right) \leq 2\exp\left(-\frac{M\epsilon^2}{2K^2}\right)$$

We have: $\mathcal{E}_M(f) = \frac{1}{M} \sum_{m=1}^M |Y_m - f(X_m)|^2$

Road map: from bounds for $\mathcal{E}_M$, to error bounds for $\widehat{f}_{\mathcal{H},M}$, in 4 steps

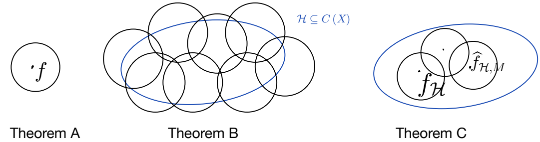
Step1: Concentration of loss for a single f

$$\mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^{M} |Y_m - f(X_m)|^2 \to \mathcal{E}_\infty(f) = \mathbb{E}[|Y - f(X)|^2]$$

Theorem (Theorem A)

Assume $|Y - f(X)| \le K$ *a.s. and* $\sigma^2 = \mathrm{Var}(Y - f(X))$*. Then,* $\forall \epsilon > 0$*,*

$$\mathbb{P}\left(|\mathcal{E}_M(f) - \mathcal{E}_\infty(f)| \ge \epsilon\right) \le 2\exp\left(-\frac{M\epsilon^2}{2\sigma^2 + \frac{2}{3}K\epsilon}\right).$$



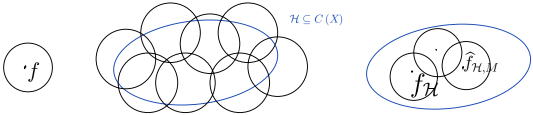
Step2: Uniform concentration of loss

Theorem (Theorem B)

Assume supp(X) is compact and let $\mathcal{H} \subset C($supp$(X))$ be compact. Assume $\sup_{f\in\mathcal{H}} |Y - f(X)| \leq K$ a.s. and $\sigma^2 = \sup_{f\in\mathcal{H}} \mathrm{Var}(Y - f(X))$. Then, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\sup_{f\in\mathcal{H}} |\mathcal{E}_M(f) - \mathcal{E}_\infty(f)| \geq \epsilon\right) \leq \mathcal{N}(\mathcal{H}, \frac{\epsilon}{8K}) 2\exp(-\frac{M\epsilon^2}{8\sigma^2 + \frac{4}{3}K\epsilon}),$$

where $\mathcal{N}(\mathcal{H}, r)$ = covering number of $\mathcal{H}$ by balls with radius r in $C($supp$(X))$.



Theorem A Theorem B Theorem C

Proof: standard argument, Finite cover + subadditivity of probability;

Step3: Bound for expected loss of estimator

$$\mathcal{E}_M(f) = \frac{1}{M}\sum_{m=1}^{M}|Y_m - f(X_m)|^2 \to \mathcal{E}_\infty(f) = \mathbb{E}[|Y - f(X)|^2]$$

$$\widehat{f}_{\mathcal{H},M} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_M(f); \quad f_{\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \mathcal{E}_\infty(f).$$

Theorem (Theorem C)

Assume: supp(X) is compact; $\mathcal{H} \subset C(\text{supp}(X))$ is compact; $\sup_{f\in\mathcal{H}} |Y - f(X)| \le K$ a.s.; $\sigma^2 = \sup_{f\in\mathcal{H}} \mathrm{Var}(Y - f(X))$. Then, $\forall \epsilon > 0$,

$$\mathbb{P}\left(\mathcal{E}_\infty(\widehat{f}_{\mathcal{H},M}) - \mathcal{E}_\infty(f_{\mathcal{H}}) > \epsilon\right) \le \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16K})2\exp(-\frac{M\epsilon^2}{32\sigma^2 + \frac{8}{3}K\epsilon}),$$

where $\mathcal{N}(\mathcal{H}, r)$ = covering number of $\mathcal{H}$ by balls with radius r in $C(\text{supp}(X))$.

Proof: By definition of $\widehat{f}_{\mathcal{H},M}$, we have $b \le 0$:

$$\mathcal{E}_\infty(\widehat{f}_{\mathcal{H},M}) - \mathcal{E}_\infty(f_{\mathcal{H}}) = \underbrace{\mathcal{E}_\infty(\widehat{f}_{\mathcal{H},M}) - \mathcal{E}_M(\widehat{f}_{\mathcal{H},M})}_{a} + \underbrace{\mathcal{E}_M(\widehat{f}_{\mathcal{H},M}) - \mathcal{E}_M(\widehat{f}_{\mathcal{H}})}_{b} + \underbrace{\mathcal{E}_M(\widehat{f}_{\mathcal{H}}) - \mathcal{E}_\infty(f_{\mathcal{H}})}_{c}.$$

$\mathbb{P}(a + b + c > \epsilon) \le \mathbb{P}(a + c > \epsilon) \le \mathbb{P}(a > \epsilon/2) + \mathbb{P}(c > \epsilon/2)$ and apply Theorem B.

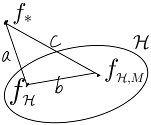
Step4: Sampling error in estimator

Theorem (Sampling error)

Assume supp(X) *compact,* $\mathcal{H} \subset C(\text{supp}(X))$ *compact convex;* $\sup_{f\in\mathcal{H}} |Y - f(X)| \le K$ *a.s.,* $\sigma^2 = \sup_{f\in\mathcal{H}} \text{Var}(Y - f(X))$*. Then,* $\forall \epsilon > 0$*,*

$$\boxed{\mathbb{P}\left(\|\widehat{f}_{\mathcal{H},M} - f_{\mathcal{H}}\|_2^2 \ge \epsilon\right) \le \mathcal{N}(\mathcal{H}, \frac{\epsilon}{16K})2\exp(-\frac{M\epsilon^2}{32\sigma^2 + \frac{8}{3}K\epsilon}),}$$

where $\mathcal{N}(\mathcal{H}, r)$ = *covering number of* $\mathcal{H}$ *by balls with radius* r *in* $C(\text{supp}(X))$*.*



- $\mathcal{E}_\infty(f) = \|f - f_*\|_2^2 + Const$
- Convexity of $\mathcal{H}$ (obtuse): $a^2 + b^2 \le c^2$

$$\Rightarrow b^2 \le c^2 - a^2$$

$$\|\widehat{f}_{\mathcal{H},M} - f_{\mathcal{H}}\|_2^2 \le \mathcal{E}_\infty(\widehat{f}_{\mathcal{H},M}) - \mathcal{E}_\infty(f_{\mathcal{H}})$$

Minimax rate: upper bound from concentration

Total error = approximation error + sampling error

$$\mathbb{E}[\|\widehat{f}_{\mathcal{H}_n,M} - f_*\|_2^2] \leq \underbrace{2\|f_{\mathcal{H}_n} - f_*\|_2^2}_{\text{Bias}} + \underbrace{2\mathbb{E}[\|\widehat{f}_{\mathcal{H}_n,M} - f_{\mathcal{H}_n}\|_2^2]}_{\text{Variance}}$$

- A bias–variance tradeoff
- Variance:
  - Covering number $\mathcal{N}(B_R, \epsilon) \leq C(\frac{R}{\epsilon})^n$
  - $\mathbb{E}[|X|] = \int_0^\infty \mathbb{P}(|X| \geq \epsilon)\, d\epsilon \leq \int_0^a d\epsilon + \int_a^\infty \mathbb{P}(|X| \geq \epsilon)\, d\epsilon \text{``} \approx \text{''} \mathrm{O}(n/M)$
- Assume bias: $\|f_{\mathcal{H}_n} - f_*\|_2^2 = \mathrm{O}(n^{-s})$:

$$C_1 \frac{n}{M} + C_2 n^{-s} = g(n) \to n_M \approx M^{\frac{1}{2s+1}}, \quad \mathbb{E}[\|\widehat{f}_{\mathcal{H}_{n_M},M} - f_*\|_2^2] \text{``} \leq \text{''} C(\frac{1}{M})^{\frac{2s}{2s+1}}$$

$$\mathbb{E}[\|\widehat{f}_{\mathcal{H}_{n_M},M} - f_*\|_2^2] \asymp C(\frac{\log M}{M})^{\frac{2s}{2s+1}}, \text{ with } n_M = (\frac{M}{\log M})^{\frac{1}{2s+1}}$$

In general: upper bound rate $\frac{2s}{2s+d}$ for $\mathbb{R}^d$-valued X.

Minimax rate: lower bound via hypothesis testing

A.B. Tsybakov. Introduction to nonparametric estimation. Springer 2008.

[To revisit in Lec3.]

- Lower bound:

$$\liminf_{M\to\infty} \inf_{\widehat{f}_M} \sup_{f\in\mathcal{C}(R,s)} \mathbb{E}_f[(M)^{\frac{2s}{2s+1}} \|\widehat{f}_M - f\|_2^2] \geq c_0 > 0.$$

- Upper bound Tsy08: Theorem 1.9,p55

$$\limsup_{M\to\infty} \sup_{f\in\mathcal{C}(R,s)} \mathbb{E}_f[(M)^{\frac{2s}{2s+1}} \|\widehat{f}_M - f\|_2^2] \leq c_1.$$

Outline

1. An overview with examples

2. Nonparametric regression and main results

3. Classical learning theory

4. Applying classical learning theory to IPS

Function space and identifiability

Learning kernels in IPS: $\dot{X}_t^i = -\frac{1}{N}\sum_{j=1}^N \phi(|X_t^i - X_t^j|)\frac{X_t^i - X_t^j}{|X_t^i - X_t^j|}$

$$
\begin{aligned}
\mathcal{E}_M(\phi) &= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt \\
&= c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.
\end{aligned}
$$

$$
\nabla \mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i \phi_i
$$

- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1} b_{n,M}$ stable?
- Identifiability of ϕ_{true}?
- Convergence of $\phi_{n,M}$? Minimax rate $\mathbb{E}\|\phi_{n_M} - \phi_{true}\|^2 \sim M^{-\frac{2s}{2s+1}}$?

Function space and identifiability

Learning kernels in IPS: $\dot{X}_t^i = -\frac{1}{N}\sum_{j=1}^N \phi(|X_t^i - X_t^j|)\frac{X_t^i - X_t^j}{|X_t^i - X_t^j|}$

$$\mathcal{E}_M(\phi) = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t^{(m)} - R_\phi(\mathbf{X}_t^{(m)})\right|^2 dt$$
$$= c^\top A_{n,M} c - 2c^\top b_{n,M} + Const.$$

$$\nabla \mathcal{E}_M = 0 \Rightarrow \quad \widehat{c} = A_{n,M}^{-1} b_{n,M} \Rightarrow \widehat{\phi}_{n,M} = \sum_i \widehat{c}_i \phi_i$$

- How to choose $\mathcal{H}_n := \mathrm{span}\{\phi_i\}_{i=1}^n$?
- $A_{n,M}^{-1}$ exists? $A_{n,M}^{-1} b_{n,M}$ stable?
- Identifiability of ϕ_{true}?
- Convergence of $\phi_{n,M}$? Minimax rate $\mathbb{E}\|\phi_{n_M} - \phi_{true}\|^2 \sim M^{-\frac{2s}{2s+1}}$?

- Exploration measure: $\rho \sim \{|X_t^i - X_t^j|\}_{i,j,t}$
- Function space: L^2_ρ
- $A_{n,M}^{-1}$ in large samle limit:

$$A_{n,\infty}(i,j) = \frac{1}{T}\int_0^T \mathbb{E}[\langle R_{\phi_i}(\mathbf{X}_t), R_{\phi_j}(\mathbf{X}_t)\rangle]dt$$
$$= \langle\langle \phi_i, \phi_j \rangle\rangle$$

Coercivity condition (CC):

$$\langle\langle \phi, \phi \rangle\rangle \geq c_{\mathcal{H}} \|\phi\|_2^2$$

- $\nabla^2 \mathcal{E}_\infty(\phi) \geq c_{\mathcal{H}} I$

Controlling estimator error by loss error: for $\mathcal{H}$ convex,

$$\boxed{\mathcal{E}_\infty(\phi) - \mathcal{E}_\infty(\phi_{\mathcal{H}}) \geq c_{\mathcal{H}} \|\phi - \phi_{\mathcal{H}}\|^2, \forall \phi \in \mathcal{H}}$$

Proof: 1. Since $\langle\langle \phi, \psi \rangle\rangle := \frac{1}{T}\int_0^T \mathbb{E}[\langle R_\phi(\mathbf{X}_t), R_\psi(\mathbf{X}_t)\rangle]dt$ and $\dot{\mathbf{X}}_t = R_{\phi_*}(\mathbf{X}_t)$:

$$\mathcal{E}_\infty(\phi) = \mathbb{E}\frac{1}{T}\int_0^T \left|\dot{\mathbf{X}}_t - R_\phi(\mathbf{X}_t)\right|^2 dt = \langle\langle \phi - \phi_* \rangle\rangle^2$$

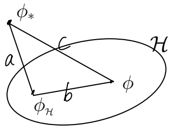
2. The obtuse inequality ($c^2 - b^2 \geq a^2$) for the bilinear form:

$$\mathcal{E}_\infty(\phi) - \mathcal{E}_\infty(\phi_{\mathcal{H}}) = \langle\langle \phi - \phi_* \rangle\rangle^2 - \langle\langle \phi_{\mathcal{H}} - \phi_* \rangle\rangle^2$$

$$\Big[= \langle\langle \phi + \phi_{\mathcal{H}} - 2\phi_*, \phi - \phi_{\mathcal{H}} \rangle\rangle \quad \text{(i.e., } |x|^2 - |y|^2 = \langle x+y, x-y\rangle \text{)}$$

$$= \langle\langle \phi - \phi_{\mathcal{H}} \rangle\rangle^2 + 2\langle\langle \phi_{\mathcal{H}} - \phi_*, \phi - \phi_{\mathcal{H}} \rangle\rangle \Big]$$

$$\geq \langle\langle \phi - \phi_{\mathcal{H}} \rangle\rangle^2 \geq c_{\mathcal{H}} \|\phi - \phi_{\mathcal{H}}\|_2^2 \quad \text{(by Coercivity)}$$



Here $\langle\langle \phi_{\mathcal{H}} - \phi_*, \phi - \phi_{\mathcal{H}} \rangle\rangle \geq 0$ by convexity of $\mathcal{H}$: $\forall t \in [0,1]$, $t\phi + (1-t)\phi_{\mathcal{H}} \in \mathcal{H}$.

$$0 \leq \mathcal{E}_\infty(t\phi + (1-t)\phi_{\mathcal{H}}) - \mathcal{E}_\infty(\phi_{\mathcal{H}}) = \langle\langle t\phi + (1-t)\phi_{\mathcal{H}} - \phi_* \rangle\rangle^2 - \langle\langle \phi_{\mathcal{H}} - \phi_* \rangle\rangle^2$$

$$= \langle\langle t\phi + (1-t)\phi_{\mathcal{H}} + \phi_{\mathcal{H}} - \phi_*, t\phi + (1-t)\phi_{\mathcal{H}} - \phi_{\mathcal{H}} \rangle\rangle$$

$$= t\langle\langle t(\phi - \phi_{\mathcal{H}}) + 2(\phi_{\mathcal{H}} - \phi_*), \phi - \phi_{\mathcal{H}} \rangle\rangle \quad \text{(send } t \to 0\text{)}$$

Main result [Theorem 6, LMT21-jmlr]:
Assuming the coercivity condition, and $\mathcal{H}$ convex+compact in $C(\mathrm{supp}(X))$. Set $n_M = (\frac{M}{\log M})^{\frac{1}{2s+1}}$, then,

$$\mathbb{E}[\|\widehat{\phi}_{n_M} - \phi_{true}\|^2_{L^2_\rho}] \leq C c_{\mathcal{H}}^{-1} (\frac{\log M}{M})^{\frac{2s}{2s+1}}.$$