Optimally tackling covariate shift in RKHS-based nonparametric regression

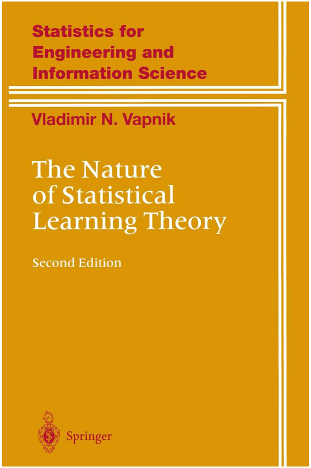
Cong Ma
Department of Statistics, UChicago

IDEAL Annual Meeting, June 2023

Reese Pathak
UC Berkeley

Martin Wainwright
MIT

Success of machine learning



Core assumption: $P_{\text{train}} = Q_{\text{test}}$

Our focus: covariate shift

$$P_{\text{train}}(X) \neq Q_{\text{test}}(X), \quad \text{while} \quad P_{\text{train}}(Y \mid X) = Q_{\text{test}}(Y \mid X)$$

due to variability in medical equipment, scanning protocols, subject populations

Key questions

- What is the statistical limit of estimation in the presence of covariate shift?
  And how does this limit depend on the "amount" of covariate shift?
- Is nonparametric least-squares estimation still optimal under covariate shift?
  If not, what is the optimal way of tackling covariate shift?

Problem setup

Nonparametric regression under covariate shift

- In standard nonparametric regression, one observes n random pairs $\{x_i, y_i\}_{i=1}^n$, where $x_i \sim P$, and

  $$y_i = f^\star(x_i) + w_i \qquad \text{with} \quad w_i \sim \mathcal{N}(0, \sigma^2)$$

  We measure performance of estimator $\widehat{f}$ by its $L^2(P)$-error:

  $$\|\widehat{f} - f^\star\|_P^2 \coloneqq \int_{\mathcal{X}} \big(\widehat{f}(x) - f^\star(x)\big)^2 p(x) dx$$

- Under covariate shift, however, our goal is to find an estimator $\widehat{f}$ whose $L^2(Q)$-error is small, where target distribution Q is different from source distribution P

Reproducing kernel Hilbert spaces (RKHSs)

- We assume throughout that $f^\star$ lies in some RKHS $\mathcal{H}$ in $L^2(Q)$
- Eigen-decomposition of kernel $\mathscr{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$:

$$\mathscr{K}(x, x') := \sum_{j=1}^{\infty} \mu_j \phi_j(x) \phi_j(x')$$

with $\{\mu_j\}_{j \geq 1}$ sequence of non-negative eigenvalues, and $\{\phi_j\}_{j \geq 1}$ orthonormal basis of $L^2(Q)$

- Hilbert norm (measure of smoothness):

$$\|f\|_{\mathcal{H}}^2 = \sum_{j=1}^{\infty} \theta_j^2 / \mu_j, \quad \text{where } \theta_j := \int_{\mathcal{X}} f(x) \phi_j(x) q(x) dx$$

- Parametrization of $\mathcal{H}$:

$$\mathcal{H} := \left\{ f = \sum_{j=1}^{\infty} \theta_j \phi_j \;\middle|\; \sum_{j=1}^{\infty} \theta_j^2 / \mu_j < \infty \right\}$$

We assume throughout that $\sup_{x \in \mathcal{X}} \mathscr{K}(x, x) \leq \kappa^2$

Examples of RKHSs

- Linear kernels: $\mathscr{K}(x, x') = \langle x, x' \rangle$ with $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{H}$ all linear functions
- Polynomial kernels: $\mathscr{K}(x, x') = (1 + \langle x, x' \rangle)^m$ with $\mathcal{X} = \mathbb{R}^d$, and $\mathcal{H}$ being polynomials of degree m or less
- First-order Sobolev space: $\mathscr{K}(x, x') = \min\{x, x'\}$ with $\mathcal{X} = [0, 1]$, and

$$\mathcal{H} = \left\{ f : [0,1] \to \mathbb{R} \mid f(0) = 0, \int_0^1 |f'(x)|^2 \mathrm{d}x < \infty \right\}$$

Family of source-target pairs

Discrepancy between $L_2(P)$ and $L_2(Q)$ norms are controlled by *likelihood ratios* (LRs)

$$\rho(x) := \frac{q(x)}{p(x)}$$

Family of source-target pairs

Discrepancy between $L_2(P)$ and $L_2(Q)$ norms are controlled by *likelihood ratios* (LRs)

$$\rho(x) := \frac{q(x)}{p(x)}$$

We focus on two broad families of covariate shift pairs (P, Q):

- Uniformly B-bounded families: $\sup_{x \in \mathcal{X}} \rho(x) \leq B$, where $B \geq 1$

  $B = 1$ recovers no-covariate-shift case

- χ^2-bounded families: $\mathbb{E}_{X \sim P}[\rho^2(X)] \leq V^2$ for some $V^2 \geq 1$

  more general than (1), and related to $\chi^2(Q||P) := \mathbb{E}_{X \sim P}[\rho^2(X)] - 1$

Uniformly B-bounded likelihood ratios

Upper bounds

A naive kernel ridge regression estimator (KRR):

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{H}} \quad \left\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \right\}$$

Theorem 1 (Ma, Pathak, Wainwright, 2022)

Assume B-bounded likelihood ratios and κ-uniformly bounded kernel. For any $\lambda \geq 10\kappa^2/n$, w.h.p. KRR $\widehat{f}_\lambda$ satisfies the bound

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \lesssim \underbrace{\lambda B \|f^\star\|_{\mathcal{H}}^2}_{\mathbf{b}_\lambda^2(B)} + \underbrace{\frac{\sigma^2 B \log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}}_{\mathbf{v}_\lambda(B)}$$

Bias-variance trade-off

Upper bound of KRR:

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \lesssim \underbrace{\lambda B \|f^\star\|_{\mathcal{H}}^2}_{\mathbf{b}_\lambda^2(B)} + \underbrace{\frac{\sigma^2 B \log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}}_{\mathbf{v}_\lambda(B)}$$

- Bias $\lambda B \|f^\star\|_{\mathcal{H}}^2$: increase as λ increases
- Variance: $\frac{\sigma^2 B \log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}$: decrease as λ increases

Bias-variance trade-off

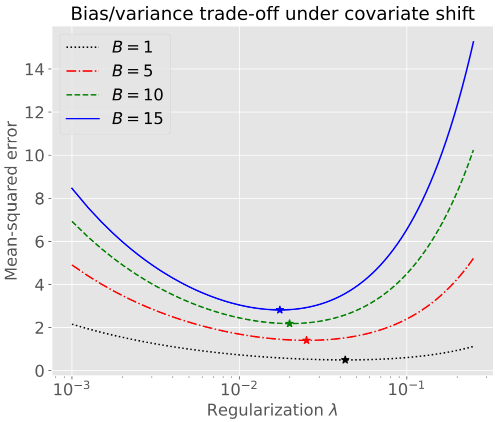
Upper bound of KRR:

$$\|\widehat{f}_\lambda - f^\star\|_Q^2 \lesssim \underbrace{\lambda B \|f^\star\|_{\mathcal{H}}^2}_{\mathbf{b}_\lambda^2(B)} + \underbrace{\frac{\sigma^2 B \log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}}_{\mathbf{v}_\lambda(B)}$$

- Bias $\lambda B \|f^\star\|_{\mathcal{H}}^2$: increase as λ increases
- Variance: $\frac{\sigma^2 B \log n}{n} \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda B}$: decrease as λ increases

Familiar! What's new?

Bias-variance trade-off



— $\mu_j = j^{-2}$, sample size $n = 8000$, and noise variance $\sigma^2 = 1$

Optimal $\lambda^*(B)$ shifts leftwards to smaller values as B is increased

Upper bounds for specific kernels

- Finite-rank kernels (i.e., $\mu_j = 0$ for $j > D$) with optimal rate $\sigma^2 B \frac{D}{n}$
- Kernels with α-decaying eigenvalues (i.e., $\mu_j \lesssim j^{-2\alpha}$) with optimal rate $(\sigma^2 B/n)^{\frac{2\alpha}{2\alpha+1}}$

Unweighted KRR is minimax optimal for these RKHSs

Sub-optimality of constrained estimator

Suppose that $\|f^\star\|_{\mathcal{H}} \le 1$. A seemingly "equivalent" estimator:

$$\widehat{f}_{\mathrm{erm}} := \arg\min_{f \in \mathcal{B}_{\mathcal{H}}(1)} \quad \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

with $\mathcal{B}_{\mathcal{H}}(1)$ denoting the ball with unit Hilbert norm

Sub-optimality of constrained estimator

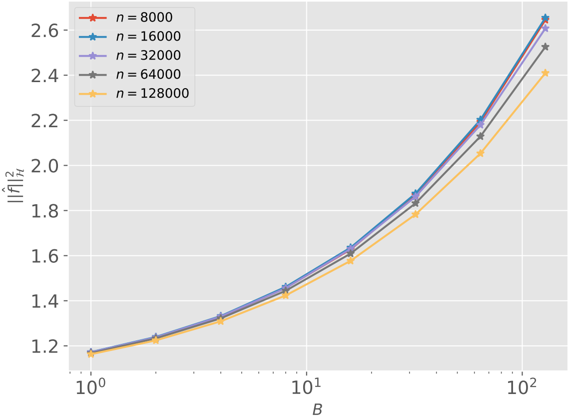
Suppose that $\|f^\star\|_{\mathcal{H}} \leq 1$. A seemingly "equivalent" estimator:

$$\widehat{f}_{\rm erm} := \arg \min_{f \in \mathcal{B}_{\mathcal{H}}(1)} \quad \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$$

with $\mathcal{B}_{\mathcal{H}}(1)$ denoting the ball with unit Hilbert norm

- Without covariate shift, constrained least-squares estimator is also rate-optimal
- However, under covariate shift, $\widehat{f}_{\rm erm}$ is provably sub-optimal. One can construct B-bounded pair (P, Q) and RKHS such that optimal rate is $(B/n)^{2/3}$, while $\mathbb{E}\Big[\|\widehat{f}_{\rm erm} - f^\star\|_Q^2\Big] \gtrsim B^3/n^2$

Intuition for failure



Key observation: $\|\widehat{f}_\lambda\|_{\mathcal{H}}^2$ increases as B increases, where $\lambda = \lambda^*(B)$

χ^2-*bounded likelihood ratios*

— going beyond uniform boundedness

A simple example

- Source distribution $P = \mathcal{N}(0, 0.9)$
- Target distribution $Q = \mathcal{N}(0, 1)$
- Unbounded likelihood ratio as $\lim_{|x| \to \infty} \rho(x) \to \infty$
- However, second moment of LRs is bounded

Unweighted KRR?

In the bounded likelihood ratio case, the key to the success of *unweighted* KRR:

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{H}} \quad \Big\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}$$

is the following nice relation

$$\boldsymbol{\Sigma}_P \succeq \tfrac{1}{B} \boldsymbol{I}$$

$\boldsymbol{\Sigma}_P := \mathbb{E}_{X \sim P}[\phi(X)\phi(X)^\top]$, and $\boldsymbol{I} = \boldsymbol{\Sigma}_Q := \mathbb{E}_{X \sim Q}[\phi(X)\phi(X)^\top]$

Unweighted KRR?

In the bounded likelihood ratio case, the key to the success of *unweighted* KRR:

$$\widehat{f}_\lambda := \arg\min_{f \in \mathcal{H}} \quad \Big\{ \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \|f\|_{\mathcal{H}}^2 \Big\}$$

is the following nice relation

$$\boldsymbol{\Sigma}_P \succeq \tfrac{1}{B} \boldsymbol{I}$$

$\boldsymbol{\Sigma}_P := \mathbb{E}_{X \sim P}[\phi(X)\phi(X)^\top]$, and $\boldsymbol{I} = \boldsymbol{\Sigma}_Q := \mathbb{E}_{X \sim Q}[\phi(X)\phi(X)^\top]$

In contrast, such a nice relationship (with B replaced by V^2) does NOT appear to hold with unbounded likelihood ratios

Likelihood-reweighted estimator

It is therefore natural to consider the likelihood-reweighted estimate

$$\arg\min_{f\in\mathcal{H}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho(x_i)(f(x_i)-y_i)^2+\lambda\|f\|_{\mathcal{H}}^2$$

- The first term is an unbiased estimate of $\mathbb{E}_Q[(Y-f(X))^2]$
- However, the variability could be huge due to multiplication by potentially *unbounded* $\rho(x)$

Therefore we consider truncated estimator

$$\widehat{f}_\lambda^{\mathrm{rw}} := \arg\min_{f\in\mathcal{H}} \quad \frac{1}{n}\sum_{i=1}^{n}\rho_{\tau_n}(x_i)(f(x_i)-y_i)^2+\lambda\|f\|_{\mathcal{H}}^2$$

Importance-reweighted estimator is near-optimal

With properly chosen λ and τ_n, $\widehat{f}_\lambda^{\mathrm{rw}}$ is optimal for a range of kernel classes including

- Finite-rank kernels with optimal rate $\frac{DV^2\sigma^2}{n}$
- Kernels with α-decaying eigenvalues with optimal rate $\left(\frac{\sigma^2 V^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$

as long as the kernel eigenfunctions are bounded $\sup_{x \in \mathcal{X}} |\phi_j(x)| \leq 1$

Conclusions and open questions

- When LRs are uniformly bounded, unweighted KRR is optimal while constrained estimator is sub-optimal
- When LRs are unbounded, likelihood reweighted KRR is optimal

Future directions:

- Prove theoretically unweighted KRR (fails to) achieve optimality
- Remove extra condition on uniformly-bounded eigen-functions

Paper:

"Optimally tackling covariate shift in RKHS-based nonparametric regression,"
C. Ma, R. Pathak, M. J. Wainwright, arXiv:2205.02986, to appear in the Annals of Statistics