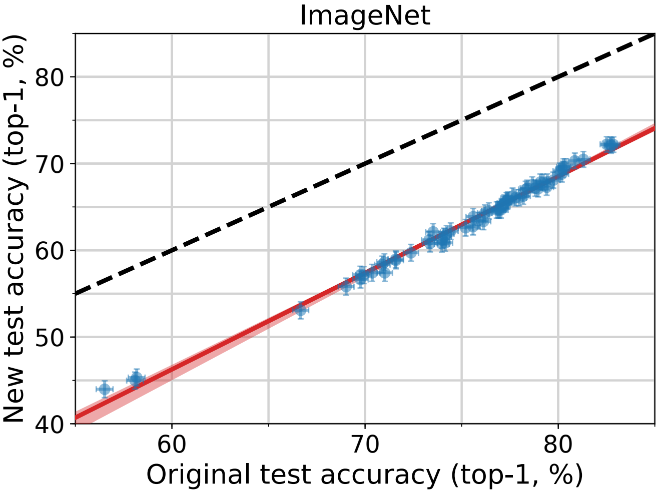
A new similarity measure for covariate shift with applications to nonparametric regression

Reese Pathak, Cong Ma, Martin J. Wainwright

ICML 2022

Challenges with distribution shift

Recht, Roelofs, Schmidt, Shankar, 2019



Regression under covariate shift

our work focuses on regression under covariate shift

observational model
we observe a dataset $\{(X_i, Y_i)\}_{i=1}^n$, where

$$Y_i = f^\star(X_i) + \xi_i, \quad i = 1, \dots, n,$$

where $f^\star = \mathbf{E}[Y \mid X = \cdot]$

covariate distribution
covariates are sampled from *source* distribution P and *target* distribution Q:

$$\begin{aligned} &\textit{source} \text{ covariates:} \quad && X_1, \dots, X_{n_P} \overset{\text{i.i.d.}}{\sim} P, \\ &\textit{target} \text{ covariates:} \quad && X_{n_P+1}, \dots, X_{n_P+n_Q} \overset{\text{i.i.d.}}{\sim} Q, \end{aligned} \qquad (n = n_P + n_Q)$$

Similarity measure

we define a measure between two distributions P, Q on metric space $(\mathscr{X}, d)$

similarity measure
for radius $h > 0$, we define

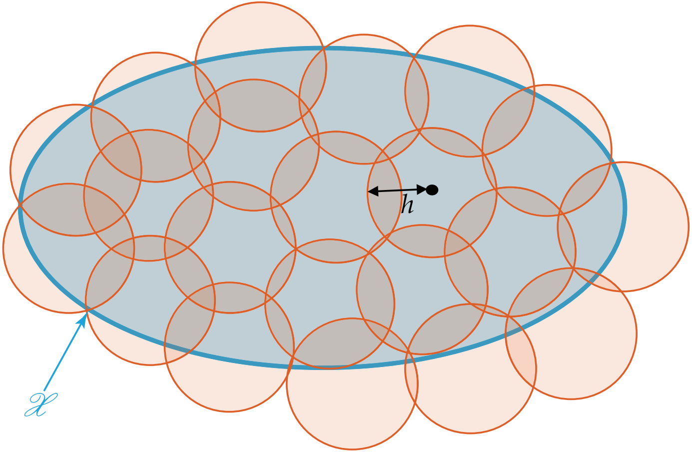
$$\rho_h(P, Q) := \int_{\mathscr{X}} \frac{1}{P(\mathsf{B}(x, h))} \, \mathrm{d}Q(x) = \mathbf{E}_{X \sim Q}\left[\frac{1}{P(\mathsf{B}(X, h))}\right]$$

above, $\mathsf{B}(x, h)$ is closed ball of radius h centered at x

- at fixed $h > 0$, absolute continuity is not required for finite similarity measure
- measure generalizes existing notions of "similarity" for pair (P, Q)
- our results use scaling of mapping $h \mapsto \rho_h(P, Q)$ in limit $h \to 0^+$

Bounds on similarity measure

we bound the similarity measure using covering numbers



covering number $N(h) :=$ minimal number of balls of radius h required to cover $\mathscr{X}$

Bounds on similarity measure

can bound similarity measure by approximating the integral over minimal covers

Proposition

Suppose that for some $h > 0$ there is $\lambda > 0$ such that the mass comparison condition

$$\lambda P(\mathsf{B}(x,h)) \geq Q(\mathsf{B}(x,h))$$

holds for all $x \in \mathscr{X}$. Then, the similarity measure satisfies

$$\rho_h(P,Q) \leq \lambda N(h/2).$$

(note λ can depend on h in claim above)

Consequences of general bound

using previous claim, can bound similarity measure in some situations

examples

- *bounded likelihood ratio:* if $Q \ll P$ and $\frac{\mathrm{d}Q}{\mathrm{d}P}(x) \leq b$ for all x, have $\rho_h(P,Q) \leq b\, N\left(\frac{h}{2}\right)$
- *transfer exponent* (Kpotufe & Martinet, 2018; 2021):
  - pair (P,Q) has (γ, C_γ)-transfer exponent if

$$P(\mathsf{B}(x,h)) \geq C_\gamma h^\gamma Q(\mathsf{B}(x,h)) \quad \text{for all } x \in \mathscr{X}, \text{ all } h > 0. \qquad (\gamma, C_\gamma) \in \mathbf{R}_+ \times (0,1]$$

  - implies similarity measure bound, $\rho_h(P,Q) \leq (h^\gamma C_\gamma)^{-1} N(h/2)$,

(note that $N(h) \lesssim h^{-k}$ as $h \to 0^+$ for compact domains $\mathscr{X} \subset \mathbf{R}^k$)

Assumptions on regression setup

recall our regression setup,

$$Y_i = f^\star(X_i) + \xi_i, \qquad \text{for } i = 1, \dots, n$$

smoothness condition
assume $\mathscr{X} = [0,1]$ and assume that $f^\star$ is L-Lipschitz,

$$f^\star \in \mathscr{F}(L) := \left\{ f : [0,1] \to \mathbf{R} \mid \left| f(x) - f(x') \right| \le L|x - x'| \text{ for any } x, x' \in [0,1] \right\}$$

noise condition
assume the noise variables satisfy (almost surely)

$$\mathbf{E}\left[\xi_i^2 \mid X_i \right] \le \sigma^2, \qquad \text{for } i = 1, \dots, n$$

Classes of covariate shifts

below are families of covariate shift instances based on the map $h \mapsto \rho_h(P, Q)$

families of covariate shifts

- we consider pairs (P, Q) for which (roughly) $\rho_h(P, Q) \lesssim h^{-\alpha}$ as $h \to 0^+$:

$$\mathscr{D}(\alpha, C) := \left\{ (P, Q) \mid \sup_{0<h\leq 1} h^{\alpha} \rho_h(P, Q) \leq C \right\} \qquad (\alpha \geq 1 \text{ and } C \geq 1)$$

- note that $\mathscr{D}(\alpha, C) \subset \mathscr{D}(\alpha', C')$ if $\alpha \leq \alpha'$ and $C \leq C'$

(some additional discussion and extensions in our full paper)

Main result: minimax upper and lower bounds

our minimax results are stated for excess prediction error under Q,

$$\|\hat{f} - f^\star\|_{L^2(Q)}^2 = \mathbf{E}_{X' \sim Q}\left[\left(\hat{f}(X') - f^\star(X')\right)^2\right].$$

Theorem

Suppose $\sigma \geq L$. Let $\alpha \geq 1, C \geq 1$. For a sufficiently large sample size, we have

$$\sup_{(P,Q) \in \mathscr{D}(\alpha,C)} \inf_{\hat{f}} \sup_{f^\star \in \mathscr{F}(L)} \mathbf{E}\|\hat{f} - f^\star\|_{L^2(Q)}^2 \asymp \left\{\left(\frac{n_P}{\sigma^2}\right)^{\frac{3}{2+\alpha}} + \left(\frac{n_Q}{\sigma^2}\right)\right\}^{-\frac{2}{3}}.$$

- when $\alpha > 1$, the worst-case rate (with no access to samples under Q) is $n^{-\frac{2}{2+\alpha}} \gg n^{-\frac{2}{3}}$
- upper bound is achieved by analyzing Nadaraya-Watson estimator under covariate shift
- lower bound is achieved by pair $(P_{\alpha,C}, Q_{\alpha,C}) \in \mathscr{D}(\alpha, C)$ that we construct

Achievable result

achievable result based on classical Nadaraya-Watson estimator

Nadaraya-Watson (NW) estimator
defined pointwise by the local average,

$$\hat{f}(x) := \frac{\sum_{i=1}^{n} Y_i \, \mathsf{1}\{X_i \in \mathsf{B}(x, h_n)\}}{\sum_{i=1}^{n} \mathsf{1}\{X_i \in \mathsf{B}(x, h_n)\}}$$

(above, $h_n > 0$ is a bandwidth parameter)

- the estimator is defined to be zero when denominator is zero
- we establish minimax upper bounds by selecting h_n as a function of $(n_P, n_Q, \sigma^2, L, \alpha, C)$

Lower bound instance

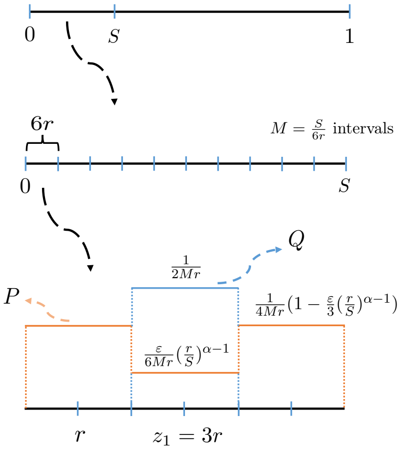


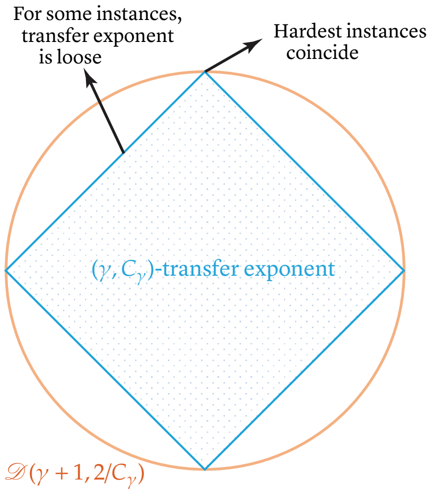
Illustration of lower bound instance

high-level overview

- we construct a hard pair $(P, Q) \in \mathscr{D}(\alpha, C)$
- we construct a hard family of regression functions within $\mathscr{F}(L)$
- we establish our minimax lower bound by combining these two pieces with Fano's inequality and packing-based arguments

Comparison to transfer exponent

introduced by Kpotufe and Martinet, 2018; 2021



our results have consequences for previously proposed notion of transfer exponent

- (P, Q) have (γ, C_γ)-transfer exponent when for all x, h

$$P(\mathsf{B}(x,h)) \geq C_\gamma h^\gamma Q(\mathsf{B}(x,h))$$

- can show if (P, Q) have (γ, C_γ)-transfer exponent, then $(P, Q) \in \mathscr{D}(\gamma+1, 2/C_\gamma)$
- consequently, can obtain upper bounds for instances with known transfer exponent

Conclusions

summary

- we introduce a similarity measure between two probability measures on the same space
- we show that this measure can be bounded easily under natural conditions
- we derive matching minimax upper and lower bounds for nonparametric regression under classes of covariate shifts that are parameterized by the scaling of this measure

additional results (not discussed)

- bounds under more general Hölder-smoothness conditions and additional classes of covariate shifts
- consequences of achievability results for bounded likelihood ratio and transfer exponent