A New Similarity Measure for Covariate Shift

Reese Pathak[1] [pathakr@berkeley.edu], Cong Ma[2], and Martin J. Wainwright[1,3]
[1]UC Berkeley Dept. of EECS, [2]University of Chicago Dept. of Statistics, [3]UC Berkeley Dept. of Statistics

Problem Setting

Regression under covariate shift

Consider a regression setting, where we observe random variables $\{(X_i, Y_i)\}_{i=1}^n$,

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

Above, $f^\star$ denotes the conditional expectation $\mathbf{E}[Y \mid X = \cdot]$. We assume we have $n = n_P + n_Q$ covariates, drawn from *source* distribution P and *target* distribution Q:

$$\textit{source} \text{ covariates:} \qquad X_1, \dots, X_{n_P} \overset{\text{i.i.d.}}{\sim} P$$

$$\textit{target} \text{ covariates:} \qquad X_{n_P+1}, \dots, X_{n_P+n_Q} \overset{\text{i.i.d.}}{\sim} Q.$$

Overview

We study the relationship between the source-target pair (P, Q) and the fundamental hardness of estimating the function $f^\star$. Specifically, we

- define a similarity measure ρ_h, based on probabilities of balls of radius $h > 0$
- relate the mapping $h \mapsto \rho_h$ to certain covering numbers of the covariate space
- characterize minimax rates over families of covariate shifts based on ρ_h

A Similarity Measure for Covariate Shift

Similarity measure

Let P, Q be two probability measures on a common metric space $(\mathscr{X}, d)$. For any radius $h > 0$, we define a *similarity measure* ρ_h as

$$\rho_h(P, Q) := \int_{\mathscr{X}} \frac{1}{P(\mathsf{B}(x, h))} \mathrm{d}Q(x),$$

where $\mathsf{B}(x, h)$ denotes the ball of radius $h > 0$ centered around x.

Properties of similarity measure

We bound the similarity measure $\rho_h(P, Q)$ via the *covering number* $N(h)$. This is the minimal number of balls of radius h required to cover $\mathscr{X}$.

Proposition. *If for some $h > 0$ there is $\lambda > 0$ such that*

$$\lambda P(\mathsf{B}(x, h)) \geq Q(\mathsf{B}(x, h)), \qquad \textit{for all } x \in \mathscr{X},$$

then we have the upper bound $\rho_h(P, Q) \leq \lambda\, N(h/2)$.

Some consequences of this result are given below.

- If $\mathscr{X} \subset \mathbf{R}^k$ has diameter D, then $\rho_h(P, Q) \leq (1 + \frac{2D}{h})^k$.
- If the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$ is uniformly bounded by b, then $\rho_h(P, Q) \leq b\, N(h/2)$.

See paper for additional examples, discussion, and the proof of this result.

Results: Minimax Upper & Lower Bounds

Assumptions

We assume $\mathscr{X} = [0, 1]$. We also assume the regression function $f^\star$ is smooth, so that some $\beta \in (0, 1]$ and $L > 0$, it lies in the Hölder class

$$\mathscr{F}(\beta, L) := \Big\{ f : [0,1] \to \mathbf{R} : |f(x) - f(x')| \leq L|x - x'|^\beta, \text{ for any } x, x' \in [0, 1] \Big\}.$$

We assume Y_i has conditional variance bounded by σ^2 almost surely.

Families of covariate shifts

We define families of covariate shifts instances—which are pairs of probability measures on $[0, 1]$. These are determined by parameters $\alpha > 0, C \geq 1$:

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

$$\mathscr{D}'(\alpha, C) := \Big\{ (P, Q) \mid \sup_{0 < h \leq 1} \big(\rho_h(Q, Q) \vee h^\alpha \rho_h(P, Q)\big) \leq C \Big\} \qquad \text{for } \alpha \in (0, 1]$$

Intuitively, these are pairs of distributions (P, Q) where the growth of the similarity measure is dominated as $\rho_h(P, Q) \lesssim h^{-\alpha}$ when $h \to 0^+$.

Main result: minimax upper & lower bounds

To estimate $f^\star$ we consider the classical Nadaraya-Watson (NW) estimator. For a parameter $h_n > 0$, it is given by

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

Below, we state matching minimax upper and lower bounds for estimating $f^\star$. Note that excess prediction error under Q is given by the norm $\|g\|_{L^2(Q)}^2 := \mathbf{E}_Q[g^2(X)]$.

Theorem. *Suppose $\sigma \geq L$. There are universal constants such that for $n_P \vee n_Q \gtrsim 1$,*
(a) for $\alpha \geq 1$ and $C \geq 1$, we have

$$\sup_{(P,Q) \in \mathscr{D}(\alpha, C)} \inf_{\hat{f}} \sup_{f^\star \in \mathscr{F}(\beta, L)} \mathbf{E}\|\hat{f} - f^\star\|_{L^2(Q)}^2 \asymp \Big\{ \Big(\frac{n_P}{\sigma^2}\Big)^{\frac{2\beta+1}{2\beta+\alpha}} + \Big(\frac{n_Q}{\sigma^2}\Big) \Big\}^{-\frac{2\beta}{2\beta+1}}, \quad \textit{and}$$

(b) for $\alpha \in (0, 1]$ and $C \geq 1$, we have

$$\sup_{(P,Q) \in \mathscr{D}'(\alpha, C)} \inf_{\hat{f}} \sup_{f^\star \in \mathscr{F}(\beta, L)} \mathbf{E}\|\hat{f} - f^\star\|_{L^2(Q)}^2 \asymp \Big\{ \Big(\frac{n_P}{\sigma^2}\Big)^{\frac{2\beta}{2\beta+\alpha}} + \Big(\frac{n_Q}{\sigma^2}\Big) \Big\}^{-1}.$$

This result summarizes Theorems 1, 2, and Corollary 1 in our full paper.

Overview of Lower Bound Argument

Proof outline

The following steps outline our construction used to prove the minimax lower bounds stated previously:

1. **Selecting a hard covariate shift pair (P, Q):** We first pick a pair $(P, Q) \in \mathscr{D}(\alpha, C)$ when $\alpha \geq 1$, or $(P, Q) \in \mathscr{D}'(\alpha, C)$ when $\alpha < 1$. The construction follows the figure on the right. The parameters $S = 6Mr$ are chosen as a function of $(\alpha, C, n_P, n_Q, \beta, \sigma, L)$ so as to vary the hardness of the instance with the problem data.
2. **Constructing hard regression functions:** We construct a family of hard regression functions $\mathscr{H}$, which have a variable number of "spikes," occurring exactly where P has low mass and Q has high mass. These spikes are constructed so as to satisfy the (β, L)-Hölder condition so that $\mathscr{H} \subset \mathscr{F}(\beta, L)$.
3. **Demonstrating hardness of instance:** Intuitively, a good estimator $\hat{f}$ of $f^\star$ must distinguish whether there is a spike (in $f^\star$) on each of the M subintervals. These regions, however, are where the likelihood ratio $\mathrm{d}Q/\mathrm{d}P$ is large. Thus, under P, we are unlikely to observe covariates there. Formally, we use a packing lower bound (Fano's method).

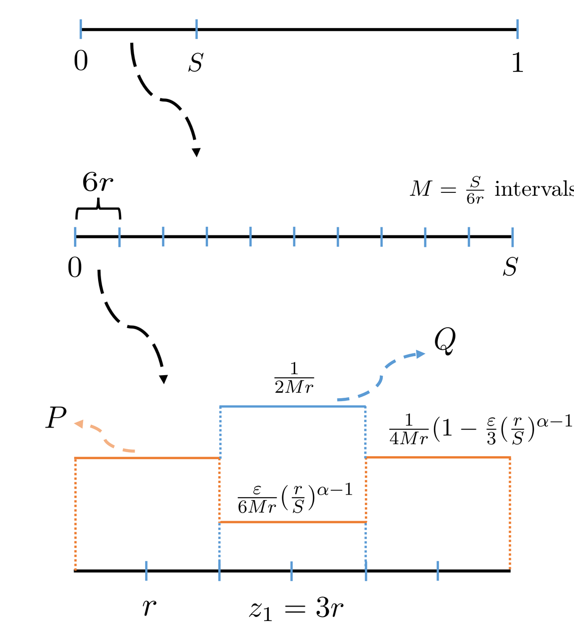
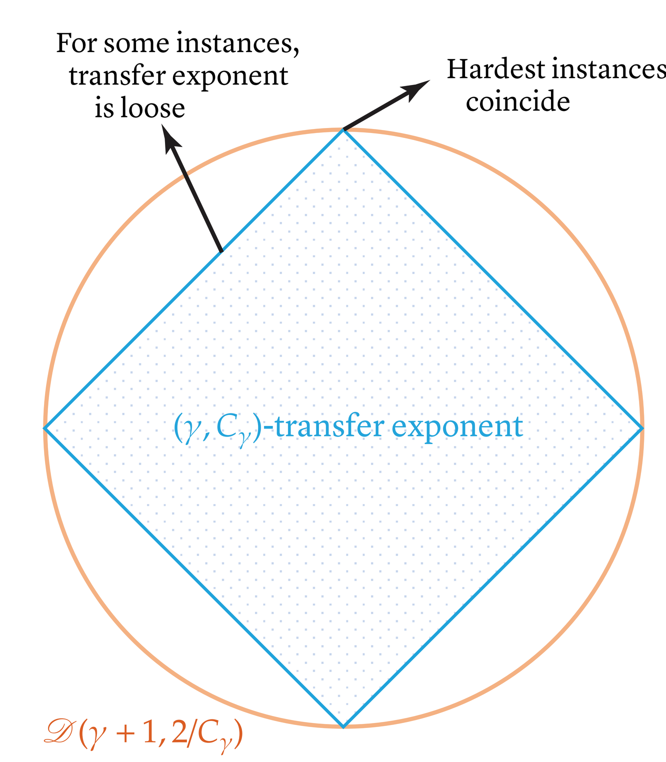


Illustration of lower bound instance

Discussion



Comparison of transfer exponent to similarity measure

Comparison to transfer exponent

Kpotufe and Martinet propose an another notion of similarity for a covariate shift pair (P, Q), defined by two parameters: $\gamma \geq 0$ and $C_\gamma \in (0, 1]$. The pair (P, Q) has (γ, C_γ)-transfer exponent if for all $h > 0$ and all $x \in \mathscr{X}$,

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

Using our proposition connecting the similarity measure with packing numbers:

$$\begin{array}{c} (P, Q) \text{ has} \\ (\gamma, C_\gamma)\text{-transfer exponent} \end{array} \implies \begin{array}{c} (P, Q) \text{ lies in} \\ \mathscr{D}(\gamma + 1, 2/C_\gamma) \end{array}$$

This implication is depicted by the figure on the left. As a result, our results imply statistical rates of convergence for our estimators when applied to covariate shift instances with known transfer exponent.

References & related work: Please see full paper (at QR code above).