Transfer Learning Through Conditional Quantile Matching

Yikun Zhang[1,†] Steven Wilkins-Reeves[2] Wesley Lee[2] Aude Hofleitner[2]

[1]Department of Statistics, University of Washington (†yikun@uw.edu) [2] Central Applied Science, Meta

Motivation

Problem: Target-domain labeled data are **scarce**; source-domain data are abundant but **shifted**.

We observe the following two categories of data:

- Target-domain data $\mathcal{D}_T = \{(X_i^{(0)}, Y_i^{(0)})\}_{i=1}^{n_0} \sim P^{(0)}$.
- K source-domain data $\mathcal{D}_S^{(k)} = \{(X_i^{(k)}, Y_i^{(k)})\}_{i=1}^{n_k} \sim P^{(k)}$ for $k = 1, \ldots, K$.

Two standard assumptions in the transfer learning literature:

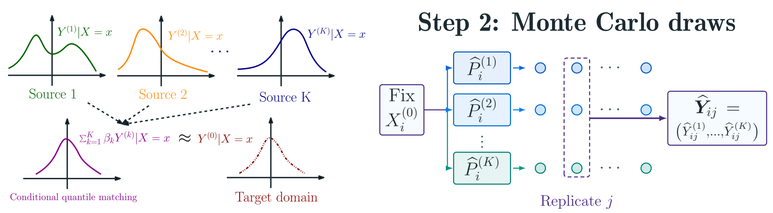
- *Covariate Shift:* $P^{(k)}(y|x) = P^{(0)}(y|x)$ but $P^{(k)}(x) \neq P^{(0)}(x)$;
- *Label Shift:* $P^{(k)}(x|y) = P^{(0)}(x|y)$ but $P^{(k)}(y) \neq P^{(0)}(y)$.

Synthetic Data Augmentation: Generate high-quality labeled data for the target domain when, for all $k = 1, \ldots, K$,

$$P^{(0)}(x) \neq P^{(k)}(x) \quad \text{and} \quad P^{(0)}(y|x) \neq P^{(k)}(y|x).$$

TLCQM Pipeline

Main Idea: Learn source conditional generators, generate synthetic responses at target covariates, then calibrate them by conditional quantile matching.



Step 1: Learn a generator $\widehat{P}^{(k)}(y|x)$ for each source domain $\mathcal{D}_S^{(k)}$.

Step 2: For target covariates $X_i^{(0)}$, draw synthetic response vectors $\widehat{\boldsymbol{Y}}_{ij} = \left(\widehat{Y}_{ij}^{(1)}, \ldots, \widehat{Y}_{ij}^{(K)}\right), j = 1, \ldots, M$ with $\widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}(y|X_i^{(0)})$.

Step 3: Compute the conditional quantile matching estimator

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{K+1}} \sum_{i=1}^{n_0} \sum_{j=1}^{M} \left[Y_{(i)}^{(0)} - \left(\boldsymbol{\beta}^T \widehat{\boldsymbol{V}}\right)_{(ij)} \right]^2, \quad \widehat{\boldsymbol{V}}_{ij} = \left(1, \widehat{\boldsymbol{Y}}_{ij}\right)^T,$$

where $Y_{(i)}^{(0)}$ and $\left(\boldsymbol{\beta}^T \widehat{\boldsymbol{V}}\right)_{(ij)}$ are the corresponding order statistics.

Step 4 (optional): Estimate $w_k(x) = \frac{dP^{(0)}(x)}{dP^{(k)}(x)}$ for covariate shift.

Step 5: Obtain the augmented target data $\mathcal{D}_T \bigcup \mathcal{D}_A$, where $\mathcal{D}_A = \bigcup_{k=1}^K \left\{ \left(X_i^{(k)}, \widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}\right) \right\}_{i=1}^{n_k}$ with $\widehat{\boldsymbol{V}}_i^{(k)} = \left(1, \widehat{Y}_i^{(1,k)}, \ldots, \widehat{Y}_i^{(K,k)}\right)$ and $\widehat{Y}_i^{(j,k)}$ as any predicted value for $Y^{(j)}$ from $\widehat{P}^{(j)}(y|X_i^{(k)})$.

Theory: Excess Risk Decomposition

With rich source data, TLCQM moves the leading ERM complexity term from n_0 to $N = \sum_{k=0}^K n_k$, while making other error dependencies explicit.

- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function, and the population risk of a prediction function $f \in \mathcal{F}$ is $R(f) := \mathbb{E}_{P^{(0)}}\left[\ell\left(Y^{(0)}, f(X^{(0)})\right)\right]$.
- $\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\left\{ \frac{1}{n} \cdot \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \sum_{i=1}^n \sigma_i \cdot f(X_i) \right| \right] \right\}$ is the Rademacher complexity of $\mathcal{F}$.
- $\mathcal{B}$ is the solution set of population-level conditional quantile matching.

Target-only Excess Risk: Let $\widehat{f}^{(0)}$ be the empirical risk minimizer (ERM) under target-only data:

$$R(\widehat{f}^{(0)}) - R(f^{(0)}) \lesssim \mathrm{Rad}_{n_0}(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{n_0}} \quad \text{w.p.} \geq 1 - \delta.$$

Excess Risk Under TLCQM: Define the transfer learning prediction function under our TLCQM framework as $\widehat{f}^{(0,tl)} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \left\{ \sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right) + \sum_{k=1}^K \sum_{i=1}^{n_k} \widehat{w}_k(X_i^{(k)}) \cdot \ell\left(\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}, f(X_i^{(k)})\right) \right\}$, leading to the bound

$$R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \lesssim \underbrace{\mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{K \log(1/\delta)}{N}}}_{\text{standard generalization error}} + \underbrace{\frac{1}{N} \sum_{k=1}^K \|\widehat{w}_k - w_k\|_1}_{\text{importance weight error}}$$

$$+ \underbrace{\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_1}_{\text{quantile matching error}} + \underbrace{\sum_{k=1}^K \|\widehat{g}^{(k)} - g^{(k)}\|_\infty}_{\text{distributional learning error}} + \underbrace{\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \left(\int_0^1 [Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}_*^T \boldsymbol{V}}(\alpha)]^2 \, d\alpha \right)^{1/2}}_{\text{transfer bias}}.$$

- Our result is **learner-agnostic**: transfer quality is separated from the downstream function class through the five explicit error terms.

Theory: Quantile Matching Error

$$\inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_*\|_1 = O\left(\sqrt{K \sum_{k=1}^K \|\widehat{g}^{(k)} - g^{(k)}\|_\infty} \right) + O\left(\frac{1}{\sqrt{M}}\right)$$
$$+ O_P\left(\sqrt{\frac{K \log\log n_0}{n_0}} + \sqrt{K} \left[\frac{\log\log n_0}{n_0} \inf_{\boldsymbol{\beta}_* \in \mathcal{B}} \int_0^1 \{Q_{Y^{(0)}}(\alpha) - Q_{\boldsymbol{\beta}_*^T \boldsymbol{V}}(\alpha)\}^2 \, d\alpha \right]^{1/4} \right).$$

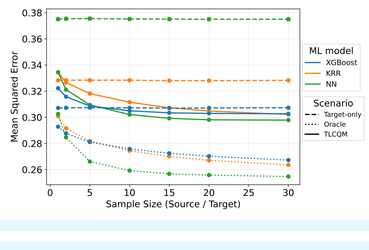
- The rate separates generator error, Monte Carlo error, target-size stochastic error, and a product of transfer bias and $O_P\left(n_0^{-\frac{1}{4}}\right)$.
- Transfer bias vanishes when the target conditional response distribution lies in the convex hull of source conditional response distributions.

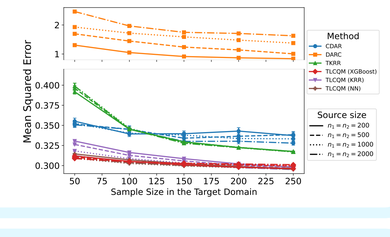
Small transfer bias improves both approximation and the convergence rate of the quantile matching objective.

Simulations: General Target Shift

We model each source-domain conditional distribution as $g^{(k)}(x, \cdot)_\# P_\eta = P^{(k)}(\cdot|x)$, and estimate $g^{(k)}$ via engression: $\widehat{g}^{(k)} \in \arg\min_{g \in \mathcal{F}} \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\frac{1}{m} \sum_{j=1}^m \left| Y_i^{(k)} - g(X_i^{(k)}, \eta_{ij}) \right| - \frac{1}{2m(m-1)} \sum_{j=1}^m \sum_{j'=1}^m \left| g(X_i^{(k)}, \eta_{ij}) - g(X_i^{(k)}, \eta_{ij'}) \right| \right]$, where P_η is a pre-specified noise distribution.

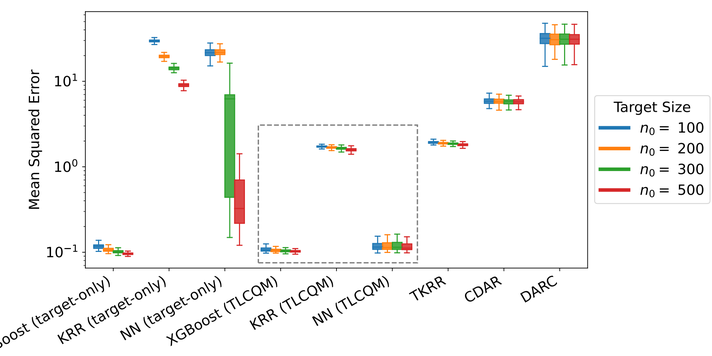
- Source domains: $Y^{(1)} = \sin\left(3\theta^T X^{(1)}\right) + 1 + \epsilon$ and $Y^{(2)} = 2\cos\left(3\theta^T X^{(2)}\right) + 1 + \epsilon$, where $\theta = \left(1, \frac{1}{2}, \ldots, \frac{1}{6}\right)^T \in \mathbb{R}^6$ and $X^{(1)}, X^{(2)} \sim \mathcal{N}(\mathbf{1}_6, \boldsymbol{I}_6), \epsilon \sim \mathcal{N}(0, 0.25)$ with $\mathbf{1}_6 = (1, \ldots, 1)^T \in \mathbb{R}^6$ and $\boldsymbol{I}_6 \in \mathbb{R}^{6 \times 6}$ being the identity matrix.
- Target domain: $Y^{(0)} = \frac{1}{3}\sin\left(3\theta^T X^{(0)}\right) - 3 + \epsilon$ with $X^{(0)} \sim \mathcal{N}(\mathbf{0}_6, 0.25 \cdot \boldsymbol{I}_6)$ and $\epsilon \sim \mathcal{N}(0, 0.25)$.





Real Data Applications

Public Apartment Data: Predict $\log$(apartment rental price) for the target state `FL` using data from source states `IL`, `OH`, and `WA`.



Monthly Active User Prediction at Meta: Predict the country-level monthly active user statistics of a target app using data from Meta's family of apps.

	Target-Only			TLCQM			TKRR	CDAR	DARC
	XGBoost	NN	KRR	XGBoost	NN	KRR			
Platform I	4.596	2.910	33.281	**1.545 (0.198)**	**1.689 (0.146)**	**1.556 (0.207)**	2.204	3.183	36.380
Platform II	2.117	2.144	27.577	**1.364 (0.034)**	**1.841 (0.015)**	**1.377 (0.034)**	2.348	1.867	10.884

- Flexible learners, especially KRR and NN, benefit most from our TLCQM framework when the target size n_0 is small.