Transfer Learning Through Conditional Quantile Matching

Yikun Zhang*
Department of Statistics, University of Chicago
NSF-Simons AI Institute for the Sky (SkAI Institute)
Chicago, IL
yikunz@uchicago.edu

Steven Wilkins-Reeves **Wesley Lee** **Aude Hofleitner**
Central Applied Science, Meta
Menlo Park, CA
{stevewr, wtlee, aude}@meta.com

Abstract

We introduce a transfer learning framework for regression that leverages heterogeneous source domains to improve predictive performance in a data-scarce target domain. Our approach fits a conditional generative model to each source domain and calibrates a shared linear combination of the generated responses via conditional quantile matching at the observed target covariates. The calibration aligns the conditional response distributions while accommodating simultaneous shifts in both the covariate and conditional response distributions. The calibrated responses then serve as pseudo-labels to augment the target data for a broad class of downstream learning tasks. Theoretically, we derive an excess risk bound for an empirical risk minimizer (ERM) trained on the augmented data, separating the benefit of an enlarged training sample from errors due to density ratio estimation, source generator estimation, conditional calibration, and irreducible transfer bias. Empirically, simulations and real-data applications show that the proposed method can improve target-domain prediction when the transfer bias is small.

1 Introduction

The rapid growth in the volume and heterogeneity of data has created new opportunities for machine learning (ML) methods to improve prediction in domains where labeled observations are scarce. Transfer learning seeks to exploit these opportunities by leveraging information from one or more related but heterogeneous source domains to improve learning in a target domain [Torrey and Shavlik, 2010, Weiss et al., 2016]. Commonly, we observe data from K source domains, $\left\{\left(X_i^{(k)}, Y_i^{(k)}\right)\right\}_{i=1}^{n_k} \sim P^{(k)}$ for $k = 1, ..., K$, together with limited data from a target domain $\left\{\left(X_i^{(0)}, Y_i^{(0)}\right)\right\}_{i=1}^{n_0} \sim P^{(0)}$, where $P^{(j)}, j = 0, 1, ..., K$ are probability distributions supported on a common product space $\mathcal{X} \times \mathcal{Y}$. The goal is to use the source data to enhance prediction in the target domain, especially when n_0 is small.

The success of transfer learning depends critically on how well the source and target distributions align. Large discrepancies can lead to negative transfer, whereby incorporating source data degrades

*Work done when the first author was an intern at Meta and affiliated with the Department of Statistics, University of Washington. Correspondence to: yikunz@uchicago.edu

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

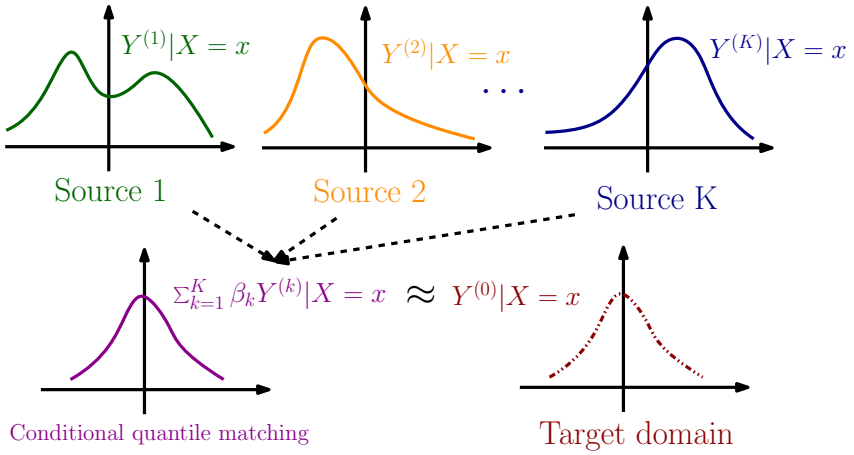


Figure 1: Illustration of conditional quantile matching at a fixed covariate value, where the intercept β_0 is omitted from the schematic.

performance in the target domain [Wang et al., 2019, Zhang et al., 2022]. To control this risk, much of the existing literature imposes structural assumptions on how source and target distributions may differ. Two widely studied assumptions are:

(i) *Covariate Shift* [Shimodaira, 2000]: $P^{(k)}(y|x) = P^{(0)}(y|x)$ but $P^{(k)}(x) \neq P^{(0)}(x)$;

(ii) *Label Shift* [Saerens et al., 2002, Nguyen et al., 2016]: $P^{(k)}(x|y) = P^{(0)}(x|y)$ but $P^{(k)}(y) \neq P^{(0)}(y)$.

Recent work has sought to relax these assumptions by introducing invariant or transformed features [Pan et al., 2010, Gong et al., 2016], latent variables [Tsai et al., 2024], or localized shift models [Wilkins-Reeves et al., 2024]. Nevertheless, empirical evidence suggests that both covariate and label shift assumptions are often violated in practice [Zhang et al., 2015, Schrouff et al., 2022], or may hold only for a subset of source domains. Moreover, existing methods that attempt to identify transferable sources typically require additional tuning parameters [Bai et al., 2024] or sample splitting of the already limited target data [Tian and Feng, 2023, Wang et al., 2023], which can further limit their effectiveness.

In this paper, we propose a novel transfer learning framework for settings in which both the covariate and conditional response distributions may vary across domains. We first learn a conditional generative model in each source domain and calibrate a shared linear combination of their generated responses by minimizing a proper scoring rule at the observed target covariates. This approach performs synthetic data augmentation for the target domain under the general distributional shift setting:

$$P^{(i)}(x) \neq P^{(j)}(x) \quad \text{and} \quad P^{(i)}(y|x) \neq P^{(j)}(y|x) \tag{1}$$

for $i \neq j$ and $i, j = 0, ..., K$. Figure 1 illustrates the central idea. At a given covariate value x, the fitted source generators produce conditional response draws, which are combined using a common coefficient vector. The key advantage of our transfer learning framework lies in its design to calibrate the conditional response distribution, rather than transferring model parameters or feature representations as commonly proposed in prior works reviewed in Section A. Moreover, the shared linear calibration implicitly regularizes the contribution of each source domain, automatically attenuating sources whose conditional responses are poorly aligned with the target. However, the calibration does not guarantee successful transfer. When the target conditional distribution is poorly approximated by the calibrated conditional family, the resulting transfer bias remains in our risk bound.

The contributions of this paper are summarized as follows.

1. **Methodology:** We introduce a general transfer learning framework for regression based on conditional quantile matching (TLCQM) in Section 2. The calibrated responses provide pseudo-labels for data augmentation and can be used with a broad range of downstream regression methods.

Algorithm 1 Data Augmentation with Conditional Quantile Matching (TLCQM)

Input: Target data $\mathcal{D}_T = \left\{\left(X_i^{(0)}, Y_i^{(0)}\right)\right\}_{i=1}^{n_0}$ and source data $\mathcal{D}_S^{(k)} = \left\{\left(X_i^{(k)}, Y_i^{(k)}\right)\right\}_{i=1}^{n_k}, k = 1, ..., K$.

Step 1: For each source domain k, fit a conditional generative model $\widehat{P}^{(k)}(y|x)$ using the source data $\mathcal{D}_S^{(k)}$.

Step 2: For each target covariate vector $X_i^{(0)}$, generate M synthetic responses independently from each fitted generator and obtain $\left\{\left(X_i^{(0)}, \widehat{\boldsymbol{Y}}_{ij}\right) : i = 1, ..., n_0, j = 1, ..., M\right\}$, where $\widehat{\boldsymbol{Y}}_{ij} = \left(\widehat{Y}_{ij}^{(1)}, ..., \widehat{Y}_{ij}^{(K)}\right)$ with $\widehat{Y}_{ij}^{(k)} \sim \widehat{P}^{(k)}(y|X_i^{(0)})$.

Step 3: Compute the conditional quantile calibration estimator $\widehat{\boldsymbol{\beta}}$ by solving (7).

Step 4: Augment the target domain with the synthetic data $\mathcal{D}_A = \bigcup_{k=1}^{K}\left\{\left(X_i^{(k)}, \widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}_i^{(k)}\right)\right\}_{i=1}^{n_k}$, with $\widehat{\boldsymbol{V}}_i^{(k)} = \left(1, \widehat{Y}_i^{(1,k)}, ..., \widehat{Y}_i^{(K,k)}\right)^T \in \mathbb{R}^{K+1}$ and $\widehat{Y}_i^{(j,k)} \sim \widehat{P}^{(j)}(y|X_i^{(k)})$ or any predicted value for $Y_i^{(j)}$ from the j-th source generator evaluated at the covariate $X_i^{(k)}$.

Step 5 (Optional): Estimate $w_k(x) = \frac{dP_X^{(0)}(x)}{dP_X^{(k)}(x)}$ for $k = 1, ..., K$ to correct for covariate shifts.

Output: The final augmented data $\mathcal{D}_F = \mathcal{D}_T \cup \mathcal{D}_A$.

2. **Theory:** We derive an excess-risk bound for ERMs trained on the augmented data in Section 3, separating standard generalization error from density ratio estimation error and conditional transfer error. We further establish a finite-sample bound for the conditional calibration estimator that quantifies the effects of source-generator error, optimization error, and irreducible transfer bias.
3. **Experiments:** Through experiments on simulated and real datasets, we demonstrate that the proposed framework can improve prediction accuracy over target-only learning and other competing transfer learning methods when transfer bias is small; see Section 4. The code used for our experiments is available at `https://github.com/facebookresearch/TLCQM`.

1.1 Notation

Throughout this paper, we consider a continuous response variable with $\mathcal{Y} \subset \mathbb{R}$ and allow a general (measurable) covariate space $\mathcal{X}$, though our proposed framework can be readily extended to settings involving multivariate and/or discrete response variables; see Section 5 for details. We say a function $f : \mathcal{X} \to \mathbb{R}$ is in $L_2(P)$ if $\|f\|_{L_2(P)} := \left[\int_{\mathcal{X}} |f(x)|^2 dP(x)\right]^{\frac{1}{2}} < \infty$ for a probability measure P. We adopt standard asymptotic notation: for deterministic sequences h_n and g_n with $g_n > 0$, we write $h_n = O(g_n)$ if $\frac{|h_n|}{g_n} \leq C$ for some absolute constant $C > 0$ and all sufficiently large n, and $h_n = o(g_n)$ if $\frac{|h_n|}{g_n} \to 0$ as $n \to \infty$. For a random sequence X_n and a positive deterministic sequence h_n, $X_n = o_P(h_n)$ means that $\frac{X_n}{h_n}$ converges to 0 in probability, while $X_n = O_P(h_n)$ indicates that $\frac{X_n}{h_n}$ is bounded in probability as $n \to \infty$. We further write $a_n \lesssim b_n$ (or equivalently $b_n \gtrsim a_n$) if $a_n \leq Cb_n$ for some constant $C > 0$ and all sufficiently large n. When both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$ hold, we denote the asymptotic equivalence by $a_n \asymp b_n$.

2 Proposed Framework

Algorithm 1 summarizes the proposed TLCQM framework. The procedure fits a conditional generative model in each source domain, calibrates a shared linear combination of the generated responses at the observed target covariates, and uses the calibrated synthetic responses to augment the target sample. Steps 1 and 3 are central to the methodology and are detailed in Section 2.1 and Section 2.2, respectively.

2.1 Learning Conditional Generative Models via Engression

Depending on data availability and problem structure, a variety of conditional generative models can be employed in Step 1 of Algorithm 1. Here, we consider a neural network-based distributional regression method known as "engression" [Shen and Meinshausen, 2025] as a default choice for its simplicity, flexibility, and empirical effectiveness. Specifically, for each source domain k, we model the conditional distribution $P^{(k)}(y|x)$ through a measurable function $g^{(k)} : \mathcal{X} \times \mathcal{E} \to \mathcal{Y}$, which maps covariates x and pre-specified noise vectors $\eta \sim P_\eta$ (*e.g.*, Gaussian or uniform) to responses. The induced pushforward measure satisfies $g^{(k)}(x, \cdot)_{\#} P_\eta = P^{(k)}(\cdot|x)$. Under the engression framework, the population-level estimator is obtained by solving

$$g^{(k)} \in \underset{g \in \mathcal{G}}{\arg\min}\, \mathbb{E}_{(X,Y,\eta,\eta') \sim P^{(k)} \times P_\eta \times P_\eta} \left[|Y - g(X,\eta)| - \frac{1}{2} |g(X,\eta) - g(X,\eta')| \right] \tag{2}$$

over a generator class $\mathcal{G}$ that contains a representation of the true conditional law, where η, η' are independent draws from P_η. The first term in (2) corresponds to an L_1-risk rewarding agreement with observed responses. The second negative term in (2) promotes variability in the generated responses across noise realizations, preventing collapse to a deterministic predictor and enabling $g^{(k)}$ to capture the conditional distribution $P^{(k)}(y|x)$. In practice, we draw $m \geq 2$ independent and identically distributed (i.i.d.) samples $\eta_{ij}, j = 1, ..., m$ for each observation $\left(X_i^{(k)}, Y_i^{(k)}\right)$ and solve the empirical optimization problem

$$\widehat{g}^{(k)} \in \underset{g \in \mathcal{G}}{\arg\min} \frac{1}{n_k} \sum_{i=1}^{n_k} \left[\frac{1}{m} \sum_{j=1}^{m} \left| Y_i^{(k)} - g(X_i^{(k)}, \eta_{ij}) \right| - \frac{1}{2m(m-1)} \sum_{j=1}^{m} \sum_{j'=1}^{m} \left| g(X_i^{(k)}, \eta_{ij}) - g(X_i^{(k)}, \eta_{ij'}) \right| \right].$$

Following Step 2 of Algorithm 1, synthetic responses are then generated at target covariates via $\widehat{Y}_{ij}^{(k)} = \widehat{g}^{(k)}(X_i^{(0)}, \eta_{ij}^{(k)})$, where $\left\{\eta_{ij}^{(k)}\right\}_{j=1}^{M}$ are i.i.d. noise vectors from the same pre-specified distribution and $M \geq 2$ controls the Monte Carlo accuracy of the calibration objective.

2.2 Conditional Quantile Matching

For $x \in \mathcal{X}$, let $\boldsymbol{V}(x, \boldsymbol{\eta}) = \left(1, g^{(1)}(x, \eta_1), ..., g^{(K)}(x, \eta_K)\right)^T \in \mathbb{R}^{K+1}$, where $\eta_1, ..., \eta_K \sim P_\eta$ are mutually independent. We also denote the conditional cumulative distribution function (CDF) of $\boldsymbol{\beta}^T \boldsymbol{V}(X, \boldsymbol{\eta})$ given $X = x$ by $G_{\boldsymbol{\beta}}(\cdot|x)$ and the target conditional CDF by $F^{(0)}(\cdot|x)$. Here, we adopt the linear form $\boldsymbol{\beta}^T \boldsymbol{V}(x, \boldsymbol{\eta})$ for two major reasons. First, it is computationally more efficient than general optimal transport-based distributional alignment methods [Chernozhukov et al., 2017]. Second, the linear structure in (3) implicitly regularizes the contribution of each source domain and mitigates negative transfer.

At the population level, our conditional quantile calibration parameter $\boldsymbol{\beta}_* \in \mathbb{R}^{K+1}$ satisfies

$$\boldsymbol{\beta}_* \in \underset{\boldsymbol{\beta} \in \Theta}{\arg\min}\, \mathcal{C}(\boldsymbol{\beta}) := \mathbb{E}\left[\mathrm{CRPS}\left(G_{\boldsymbol{\beta}}(\cdot|X^{(0)}), Y^{(0)} \right) \right]. \tag{3}$$

Here, we optimize over a compact set $\Theta = \{\boldsymbol{\beta} \in \mathbb{R}^{K+1} : \|\boldsymbol{\beta}\|_1 \leq B_\beta\}$, which can be enforced by adding a L_1 regularization term to the objective (3) in practice. Additionally,

$$\mathrm{CRPS}(G, y) = \int_{\mathbb{R}} \left[G(t) - \mathbb{1}\{y \leq t\} \right]^2 dt = 2 \int_0^1 \rho_\alpha \left(y - Q_G(\alpha) \right) d\alpha \tag{4}$$

is the continuous ranked probability score (CRPS; Matheson and Winkler 1976, Hersbach 2000), where $\rho_\alpha(u) = u\left[\alpha - \mathbb{1}\{u < 0\}\right]$ is the quantile loss and $Q_G(\alpha)$ is the α-quantile of the CDF G. The second expression of CRPS in (4) illustrates that (3) minimizes the *conditional quantile loss* of $\boldsymbol{\beta}^T \boldsymbol{V}(x, \boldsymbol{\eta})$ with $Y^{(0)}$ given $X = x$ integrated over $\alpha \in [0, 1]$ and averaged over target covariates.

When G has a finite first moment, CRPS has the following more computationally tractable representation as:

$$\mathrm{CRPS}\,(G, y) = \mathbb{E}_{Z\sim G}|y - Z| - \frac{1}{2}\mathbb{E}_{Z,Z'\overset{\text{ind}}{\sim}G}|Z - Z'|, \tag{5}$$

where Z and Z' are independent random variables with CDF as G; see Lemma 2.2 in Baringhaus and Franz [2004]. As proved by Gneiting and Raftery [2007], CRPS is a strictly proper score. In particular, the excess of $\mathcal{C}(\boldsymbol{\beta})$ in (3) over the oracle minimum $\mathcal{C}_0 = \mathbb{E}\left\{\mathrm{CRPS}\left[F^{(0)}(\cdot|X^{(0)}), Y^{(0)}\right]\right\}$ is given by

$$\mathcal{C}(\boldsymbol{\beta}) - \mathcal{C}_0 = \mathbb{E}_{X^{(0)}}\left\{\int_{\mathbb{R}}\left[F^{(0)}(t|X^{(0)}) - G_{\boldsymbol{\beta}}(t|X^{(0)})\right]^2 dt\right\} =: \Delta(\boldsymbol{\beta}). \tag{6}$$

According to (5), we compute the sample conditional quantile calibration estimator in Step 3 of Algorithm 1 as:

$$\widehat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\Theta} \widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}) \qquad \text{with}$$
$$\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}) := \frac{1}{n_0}\sum_{i=1}^{n_0}\left\{\frac{1}{M}\sum_{j=1}^{M}\left|Y_i^{(0)} - \boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij}\right| - \frac{1}{2M(M-1)}\sum_{j,j'=1}^{M}\left|\boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij'}\right|\right\} \tag{7}$$

where $\widehat{\boldsymbol{V}}_{ij} = \left(1, \widehat{Y}_{ij}^{(1)}, ..., \widehat{Y}_{ij}^{(K)}\right)^T \in \mathbb{R}^{K+1}$ consists of generated synthetic responses $\widehat{Y}_{ij}^{(k)} = \widehat{g}^{(k)}(X_i^{(0)}, \eta_{ij}^{(k)}), j = 1, ..., M$ at each $X_i^{(0)}$ for $k = 1, ..., K$. Conditional on source generators $\widehat{g}^{(k)}, k = 1, ..., K$, fitted independently of the target observations, the objective $\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta})$ in (7) is an unbiased estimator (*i.e.*, U-statistic) of $\mathbb{E}\left[\mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X^{(0)}), Y^{(0)}\right)\right]$ for any fixed $\boldsymbol{\beta} \in \Theta$. When $\widehat{g}^{(k)}$ is consistent for $g^{(k)}$, $\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta})$ also consistently estimates $\mathcal{C}(\boldsymbol{\beta})$.

While the objective in (7) is non-smooth and generally non-convex, it is a difference of a proper closed convex function and a continuous convex function, which can be optimized by the classical difference-of-convex algorithm (DCA; Tao and An 1997, 1998). Here, we use a proximal DCA update related to the framework of Wen et al. [2018]. Writing $\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}) = \Phi(\boldsymbol{\beta}) - \Psi(\boldsymbol{\beta})$, where

$$\Phi(\boldsymbol{\beta}) = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{1}{M}\sum_{j=1}^{M}\left|Y_i^{(0)} - \boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij}\right|, \quad \Psi(\boldsymbol{\beta}) = \frac{1}{n_0}\sum_{i=1}^{n_0}\frac{1}{2M(M-1)}\sum_{j,j'=1}^{M}\left|\boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij} - \boldsymbol{\beta}^T\widehat{\boldsymbol{V}}_{ij'}\right|$$

are both convex, the proximal DCA chooses a subgradient $\boldsymbol{s}^{(r)} \in \partial\Psi(\boldsymbol{\beta}^{(r)})$ and updates

$$\boldsymbol{\beta}^{(r+1)} \in \arg\min_{\boldsymbol{\beta}\in\Theta}\left\{\Phi(\boldsymbol{\beta}) - \boldsymbol{s}^{(r)T}\boldsymbol{\beta} + \frac{\gamma}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{(r)}\|_2^2\right\}, \qquad \text{for some } \gamma > 0. \tag{8}$$

The convergence results of this proximal DCA update (8) have been studied in Wen et al. [2018].

For squared-error prediction, we recommend using the calibrated conditional mean rather than an individual generated response as the augmented label in Step 4 of Algorithm 1. In particular, at any source covariate $X_i^{(k)}$, we replace $\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}_i^{(k)}$ with

$$\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X_i^{(k)}) = \widehat{\beta}_0 + \sum_{j=1}^{K}\widehat{\beta}_j\left[\frac{1}{M'}\sum_{\ell=1}^{M'}\widehat{g}^{(j)}(X_i^{(k)}, \eta_\ell^{(j)})\right], \tag{9}$$

where $M' > 0$ is an integer controlling the Monte Carlo accuracy of the estimated conditional mean. This construction removes unnecessary response noise from the pseudo-labels and leads to a smaller conditional transfer error; see Theorem 3.4 for details.

Remark 2.1. Let $\mu^{(0)}(x) = \mathbb{E}(Y^{(0)}|X^{(0)} = x)$ be the true regression function in the target domain and $\mu^{(k)}(x) = \mathbb{E}_\eta\left[g^{(k)}(x, \eta)\right]$. When $g^{(k)}$ is correctly specified (see Assumption 3.2 below), $\mu^{(k)}(x) = \mathbb{E}(Y^{(k)}|X^{(k)} = x)$. The conditional mean induced by $\boldsymbol{\beta}$ is written as $\mu_{\boldsymbol{\beta}}(x) = \beta_0 + \sum_{k=1}^{K}\beta_k \cdot \mu^{(k)}(x)$. If $\mu^{(0)}$ lies outside the calibrated mean class $\{\mu_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Theta\}$, no common coefficient vector can eliminate the conditional mean discrepancy. This irreducible discrepancy contributes to the transfer bias in our theoretical analysis; see Theorem 3.5 for details.

3 Theoretical Analysis

This section theoretically compares the prediction risks of ERMs trained solely on the target domain with those trained on the augmented data $\mathcal{D}_F$ produced by our TLCQM framework (Algorithm 1).

Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a symmetric loss function. The population risk of a prediction function $f \in \mathcal{F}$ on the target domain is defined as $R(f) := \mathbb{E}\left[\ell\left(Y^{(0)}, f(X^{(0)})\right)\right]$, and let $f^{(0)} \in \arg\min_{f\in\mathcal{F}} R(f)$ be the population risk minimizer over the function class $\mathcal{F}$. Then, the excess risk of any candidate function $f : \mathcal{X} \to \mathcal{Y}$ is defined as $R(f) - R(f^{(0)}) = \mathbb{E}\left[\ell\left(Y^{(0)}, f(X^{(0)})\right)\right] - \mathbb{E}\left[\ell\left(Y^{(0)}, f^{(0)}(X^{(0)})\right)\right]$.

We analyze ERMs corresponding to the following two different training strategies.

- **(Target-only ERM)** We define $\widehat{f}^{(0)} : \mathcal{X} \to \mathcal{Y}$ as:

$$\widehat{f}^{(0)} \in \arg\min_{f\in\mathcal{F}} \frac{1}{n_0} \sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right). \tag{10}$$

- **(TLCQM ERM)** Using the augmented data $\mathcal{D}_F$ from Algorithm 1, the transfer learning prediction function is defined as:

$$\widehat{f}^{(0,tl)} \in \arg\min_{f\in\mathcal{F}} \frac{1}{N} \left\{ \sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right) + \sum_{k=1}^{K} \sum_{i=1}^{n_k} \widehat{w}_k(X_i^{(k)}) \cdot \ell\left(\widehat{Z}_i^{(k)}, f(X_i^{(k)})\right) \right\}, \tag{11}$$

  where $N = \sum_{k=0}^{K} n_k$ is the combined sample size of both source and target domains, $\widehat{w}_k(x)$ is the estimated density ratio $w_k(x) = \frac{dP_X^{(0)}(x)}{dP_X^{(k)}(x)}$ for $k = 1, ..., K$, and $\widehat{Z}_i^{(k)}$ is either the Monte Carlo augmented conditional mean in (9) or an independent calibrated draw $\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}$.

The density ratio correction in (11) is not necessary when the function class $\mathcal{F}$ for the prediction task is correctly specified [Sugiyama et al., 2007]. In practice, however, incorporating importance weights improves robustness to model misspecification and often stabilizes numerical optimization.

3.1 Assumptions

To state a finite-sample result without imposing uniform entropy conditions on data-dependent nuisance estimators, we assume that the fitted generators $\widehat{g}^{(k)}, k = 1, ..., K$, calibration coefficients $\widehat{\boldsymbol{\beta}}$, and density ratio estimators $\widehat{w}_k, k = 1, ..., K$ are learned from auxiliary data independent of the observations used to train the downstream ERM. The same guarantee can be implemented without losing data efficiency through cross-fitting [Schick, 1986, Newey and Robins, 2018].

Assumption 3.1 (Basic regularity conditions)**.**

(a) On a compact interval containing $\mathcal{Y}$ and all calibrated responses used for augmentation, there exists a constant $B_\ell > 0$ such that $|\ell(y, y')| \le B_\ell$ and $|\ell(y, y') - \ell(y, y'')| \le B_\ell |y' - y''|$.

(b) $P_X^{(0)}$ is absolutely continuous with respect to each $P_X^{(k)}$, and there exists a constant $B_w > 0$ such that $0 \le w_k(x) = \frac{dP_X^{(0)}(x)}{dP_X^{(k)}(x)} \le B_w$ almost surely for all $k = 1, ..., K$.

Assumption 3.2 (Boundedness conditions for excess risk)**.** The conditional distribution $P^{(k)}(Y^{(k)}|X = x)$ is represented by $g^{(k)}(x, \eta)$ for $k = 0, 1, ..., K$, where η is an independent random vector with a pre-specified distribution. For a constant $B_g > 0$, suppose $\sup_{(x,\eta)} |g^{(k)}(x, \eta)| \le B_g$ for $k = 0, 1, ..., K$ and $\sup_{(x,\eta)} |\widehat{g}^{(k)}(x, \eta)| \le B_g$ for $k = 1, ..., K$.

Assumption 3.1(a) is a standard boundedness and Lipschitz continuity condition on both arguments of the loss function, which holds, for example, when $\mathcal{Y}$ is compact or when the loss function ℓ is truncated at a certain value $B_\ell > 0$. Assumption 3.1(b) prevents instability from extreme density ratios (or importance weights) under covariate shifts [Gretton et al., 2009, Reddi et al., 2015], so that these weights can be consistently estimated. Assumption 3.2, together with the compactness of Θ, uniformly bounds the true and fitted conditional generative models, which also resemble the compactness assumption on $\mathcal{Y}$.

3.2 Target-Only Excess Risk

For a function class $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ and independent covariates $X_1, ..., X_n$ (not necessarily identically distributed), let $\widehat{\mathrm{Rad}}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\sigma[\sup_{f\in\mathcal{F}} |\sum_{i=1}^n \sigma_i \cdot f(X_i)|]$, where σ_i's are i.i.d. Rademacher random variables, and denote $\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\left[\widehat{\mathrm{Rad}}_n(\mathcal{F})\right]$. For $n = N$, the expectation is taken over the joint distribution of the pooled target and source covariates.

Proposition 3.3. *Under Assumption 3.1(a), the excess risk of the target-only prediction function $\widehat{f}^{(0)}$ in (10) satisfies, with probability at least $1-\delta$,*

$$R(\widehat{f}^{(0)}) - R(f^{(0)}) \lesssim \mathrm{Rad}_{n_0}(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{n_0}}.$$

Proposition 3.3 is a classical result in statistical learning theory; see, *e.g.*, Theorem 26.5 in Shalev-Shwartz and Ben-David [2014]. The upper bound remains valid if $\mathrm{Rad}_{n_0}(\mathcal{F})$ is replaced by its empirical counterpart $\widehat{\mathrm{Rad}}_{n_0}(\mathcal{F})$. In many common settings, $\mathrm{Rad}_{n_0}(\mathcal{F})$ scales as $O\left(\frac{1}{\sqrt{n_0}}\right)$ up to some $\log n_0$ factor, with constants depending critically on the complexity of the functional class $\mathcal{F}$. The representative bounds are summarized in Table 2.

3.3 Excess Risk Under Our TLCQM framework

For $\boldsymbol{\beta} \in \Theta$, let $\widehat{G}_{\boldsymbol{\beta}}(\cdot|x)$ be the conditional CDF of $\boldsymbol{\beta}^T\widehat{\boldsymbol{V}}(x, \boldsymbol{\eta}) = \widehat{\beta}_0 + \sum_{k=1}^K \widehat{\beta}_k \cdot \widehat{g}^{(k)}(x, \eta_k)$, and we define its conditional CRPS error by

$$\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) := \mathbb{E}_{X^{(0)}}\left\{\int_{\mathbb{R}} \left[F^{(0)}(t|X^{(0)}) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|X^{(0)})\right]^2 dt\right\}. \tag{12}$$

Theorem 3.4. *Suppose that Assumptions 3.1 and 3.2 hold and that $\left(\widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}}\right)$ is fitted on auxiliary data independent of the observations used in (11). Assume further that $\ell(y, y') = (y - y')^2$ and $\mathcal{F}$ is a closed convex subset of $L_2(P_X^{(0)})$. Then, the transfer learning prediction function $\widehat{f}^{(0,tl)}$ in (11) under augmented responses as (9) has its excess risk satisfying that, with probability at least $1-\delta$,*

$$R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \lesssim \underbrace{\mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{N}}}_{\textit{Standard generalization error}} + \underbrace{\frac{1}{N}\sum_{k=1}^K \|\widehat{w}_k - w_k\|_1}_{\textit{Importance weight error}}$$
$$+ \underbrace{\frac{(N-n_0)^2}{N^2}\left[\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) + \frac{1}{M'}\right]}_{\textit{Conditional transfer and Monte Carlo errors}}$$

where $\|\widehat{w}_k - w_k\|_1 = \sum_{i=1}^{n_k}\left|\widehat{w}_k(X_i^{(k)}) - w_k(X_i^{(k)})\right|$ and $\delta \in (0,1)$.

If $\mathcal{F}$ is not convex, ℓ is a general loss function satisfying Assumption 3.1(a) only, and the augmented labels are independent calibrated draws $\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}_i^{(k)}$, then the last term becomes $\frac{N-n_0}{N}\sqrt{\widehat{\Delta}(\widehat{\boldsymbol{\beta}})}$.

The proof of Theorem 3.4 is in Section E.2, where the transfer-learning excess risk is decomposed into three error terms.

- **Standard generalization error:** This term mirrors the classical ERM bound but scales with the augmented sample size N rather than the small target sample size n_0. Under standard complexity scaling, it is typically smaller when $N \gg n_0$.
- **Importance weight error:** This term captures the error induced by estimating density ratios under covariate shift. When the penalized risk minimization [Nguyen et al., 2007] is applied, the rate is $O_P\left(\sum_{k=1}^K\left(n_k^{-\frac{1}{2+\tau}} + n_k^{-\frac{1}{4}}\right)\right)$ up to some $\log n_k$ factors (see Lemma 8 in Reddi et al. 2015). This term vanishes faster than the target-only excess risk under mild growth of n_k for $k = 1, ..., K$.

- **Conditional transfer error:** We further derive an upper bound for the conditional CRPS error (12) in Theorem 3.5 below. In particular,

$$\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) \leq \underbrace{\Delta(\boldsymbol{\beta}_*)}_{\textbf{Transfer bias}} + \underbrace{\sum_{k=1}^{K} \left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty}_{\textbf{Distributional learning error}} + O_P\left(\sqrt{\frac{\log n_0 \left[\Delta(\boldsymbol{\beta}_*) + \sum_{k=1}^{K} \left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty \right]}{n_0}} \right) \tag{13}$$

  up to some optimization and Monte Carlo approximation errors.

  The "**transfer bias**" term is the intrinsic approximation error of the population-level conditional quantile matching (3) relative to the target distribution. By (6), $\Delta(\boldsymbol{\beta}_*) = 0$ if there exists $\boldsymbol{\beta} \in \Theta$ such that the conditional distributions of $\boldsymbol{\beta}^T V(X, \boldsymbol{\eta}) = \beta_0 + \sum_{k=1}^{K} \beta_k g^{(k)}(X^{(0)}, \eta_k)$ equals $F^{(0)}(\cdot|X^{(0)})$ for $P_X^{(0)}$-almost every x, *i.e.*, $P^{(0)}(Y|X)$ lies in the distribution family induced by the affine span of $P^{(k)}(Y|X), k = 1, ..., K$, a setting commonly assumed in multi-source transfer learning (see Section 3 in Turrisi et al. 2022). In this case, the proposed framework achieves asymptotically unbiased transfer learning.

  The "**distributional learning error**" depends on the expected uniform prediction errors of the source generators. Under standard smoothness assumptions, the optimal nonparametric error rate is $O\left(\sum_{k=1}^{K} n_k^{-\frac{s}{d+2s}}\right)$, where s is the Hölder smoothness level of $g^{(k)}$ for $k = 1, ..., K$ [Stone, 1980, 1982]. Related rates are attainable by sieve methods [Chen and Christensen, 2013] and neural networks [Imaizumi, 2023]. Within the engression framework, when $g^{(k)}$ is quadratic in x and η, $\left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty$ converges at rate $O_P\left(n_k^{-\frac{1}{3}}\right)$ up to a logarithmic factor.

  The third term in (13) scales as $O_P\left(\sqrt{\frac{\log n_0}{n_0}}\right)$ times the square root of the sum of the transfer bias and distributional learning error. It is therefore of smaller order when both components vanish sufficiently quickly.

Finally, we emphasize that when the optimal quantile matching estimator (7) is sparse, the K-dependent terms in Theorem 3.4 scale with the number of active source domains rather than the total number of source domains.

3.4 Conditional Quantile Calibration Error

Theorem 3.5. *Let $\delta \in (0, 1)$ and $\left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty = \sup_{x \in \mathcal{X}} \mathbb{E}_\eta \left| \widehat{g}^{(k)}(x, \eta) - g^{(k)}(x, \eta) \right|$ for $k = 1, ..., K$. Suppose that Assumption 3.2 holds and that $\widehat{\boldsymbol{g}}$ is independent of the target observations used in (7). If $\widehat{\mathcal{C}}_{n_0,M}(\widehat{\boldsymbol{\beta}}) \leq \inf_{\boldsymbol{\beta} \in \Theta} \widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}) + \epsilon_{\text{opt}}$ in (7), then with probability at least $1 - \delta$,*

$$\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) - \Delta(\boldsymbol{\beta}_*) \lesssim \sum_{k=1}^{K} \left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty + \sqrt{\frac{(\log n_0 + \log(1/\delta)) \left[\Delta(\boldsymbol{\beta}_*) + \sum_{k=1}^{K} \left\| \widehat{g}^{(k)} - g^{(k)} \right\|_\infty \right]}{n_0}} + \frac{\log n_0 + \log(1/\delta)}{n_0} + \epsilon_{\text{opt}}$$

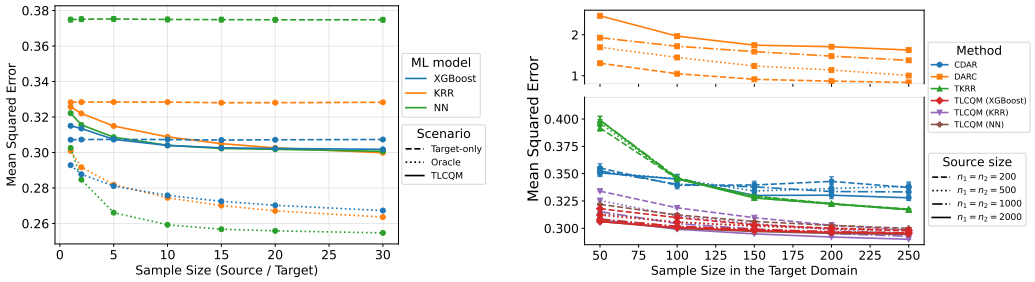
up to some Monte Carlo approximation errors of order $O\left(\frac{1}{\sqrt{M}}\right)$.

The proof of Theorem 3.5 is in Section E.3.

4 Numerical Experiments

In this section, we empirically evaluate the performance of several ML models under the proposed TLCQM framework in Algorithm 1 and compare them with target-only ML models. We also benchmark TLCQM against several competing transfer learning approaches for regression under distributional shift, including (i) transfer learning with kernel ridge regression and automatic source selection (TKRR; Wang et al. 2023), (ii) deep transfer learning for conditional shift in regression (CDAR; Liu et al. 2021), and (iii) multi-domain adaptation for regression under conditional shift (DARC; Taghiyarrenani et al. 2023). Implementation details for all methods are provided in Section B.

4.1 Simulation Studies



(a) MSE as a function of the source-to-target sample size ratio for ML models under three data scenarios.

(b) MSE as a function of the target-domain sample size for different transfer learning methods.

Figure 2: Prediction performance of different ML models under the simulated training scenarios and competing transfer learning methods.

We consider a simulation setting in which both covariate shift and concept shift are present between the source and target domains. For the two candidate source domains, $Y^{(1)} = \sin\left(3\theta^T X^{(1)}\right)+1+\epsilon^{(1)}$ and $Y^{(2)} = 2\cos\left(3\theta^T X^{(2)}\right) + 1 + \epsilon^{(2)}$, where $\theta = \left(1, \frac{1}{2}, ..., \frac{1}{5}\right)^T \in \mathbb{R}^5$, $X^{(1)}, X^{(2)} \sim \mathcal{N}(\mathbf{1}_5, \boldsymbol{I}_5)$, and $\epsilon^{(1)}, \epsilon^{(2)} \sim \mathcal{N}(0, 0.25)$. Here, $\mathbf{1}_5 = (1, ..., 1)^T \in \mathbb{R}^5$ and $\boldsymbol{I}_5 \in \mathbb{R}^{5\times 5}$ denote the vector of ones and the identity matrix, respectively. For the target domain, $Y^{(0)} = \frac{1}{3}\sin\left(3\theta^T X^{(0)}\right) - 3 + \epsilon^{(0)}$, where $X^{(0)} \sim \mathcal{N}(\mathbf{0}_5, 0.25 \cdot \boldsymbol{I}_5)$ and $\epsilon^{(0)} \sim \mathcal{N}(0, 0.25)$.

We vary the source-to-target sample size ratio $\frac{n_k}{n_0}$ from 1 to 30 and the target-domain sample size n_0 from 50 to 250 in Figure 2. To assess the effect of data augmentation by Algorithm 1, we consider three widely used ML models: XGBoost [Chen and Guestrin, 2016], kernel ridge regression (KRR), and neural network (NN). We apply them under three training scenarios: target-only training, oracle training, and augmented training with TLCQM. The oracle data are drawn directly from the target-domain distribution and have the same sample size as the combined source and target datasets.

Figure 2(a) reports the average mean squared errors (MSEs) across 1000 Monte Carlo repetitions at each source-to-target sample size ratio $\frac{n_k}{n_0}$. In this simulation setting, training on the TLCQM-augmented data generally improves prediction relative to target-only training, with larger gains when the source-to-target sample size ratio is high. The improvements are most pronounced for NN and KRR in the reported experiments. We further compare our TLCQM framework with competing transfer learning methods that directly leverage both source and target data. As shown in Figure 2(b), all ML models trained on TLCQM-augmented data outperform TKRR, CDAR, and DARC across the considered target sample sizes. Additional results for ML models trained on source-only data, naive pooling of source and target data, and simple target-only data augmentation are reported in Section C.1, together with ablation studies in Section C.2 and examination of negative transfer in Section C.3. We also conduct evaluations on a public dataset in Section D.

4.2 Case Study: Monthly Active User Prediction

We apply our proposed TLCQM framework (Algorithm 1) to a real-world task involving the prediction of monthly active users (MAUs) for peer apps within Meta. The dataset consists of country-level MAU measurements collected over a single month, covering two device platforms and five distinct apps. To alleviate the heavy-tailed distribution, we apply a logarithmic transformation $u \mapsto \log(1+u)$ to all MAU measurements in this case study. Each app is treated as a separate data domain, with approximately 200 observations corresponding to different countries or platforms. The feature vector X includes MAU statistics from other apps and a set of country-specific covariates, while the response $Y \in [0, \infty)$ represents the (log-transformed) country-level MAU for each app.

Our goal is to predict the country-level MAU for an app developed by a peer company, for which the third-party MAU estimates are available for a subset of countries. This application motivates the general shift setting (1), since both covariate and conditional country-level MAU distributions may vary across apps. To emulate this business scenario while enabling systematic evaluation, we use

	Target-Only			TLCQM			TKRR	CDAR	DARC
	XGBoost	NN	KRR	XGBoost	NN	KRR			
Platform I	9.180	5.717	21.794	**4.099 (0.014)**	**4.211 (0.012)**	**4.073 (0.014)**	6.981	139.327 (1.093)	18.220 (1.156)
Platform II	8.998	5.005	32.218	**3.785 (0.065)**	**3.840 (0.046)**	**3.767 (0.068)**	4.967	117.149 (1.194)	11.542 (0.921)

Table 1: Average MSEs of different ML models trained on target-only and TLCQM-augmented data, together with those of competing transfer learning methods, for the MAU prediction task across two device platforms. Standard errors of the MSEs based on 30 Monte Carlo experiments are reported in parenthesis, where those values smaller than 0.001 are omitted for brevity.

leave-one-app-out validation, where 4 out of the 5 apps are randomly selected as source domains with complete MAU information, and the remaining app is treated as the target domain, for which MAU statistics are observed only for the aforementioned subset of countries. The proposed TLCQM framework is then employed to augment or impute the missing MAU values for the remaining countries in the target app. To ensure the nonnegativity of the generated MAU values, we impose nonnegative constraints on the coefficients of the quantile matching estimator (7) in our proposed framework, which can be interpreted as an implicit form of regularization tailored to this data application.

The empirical results are summarized in Table 1, which reports the average MSEs for each ML model trained on data augmented by our proposed TLCQM framework. For comparison, we also include results from ML models trained on the target-only data as well as other competing transfer learning methods that leverage both source and target data, including TKRR, CDAR, and DARC that we mentioned earlier. Since our proposed Algorithm 1 involves generative model learning and synthetic data generation, their reported MSEs exhibit a small degree of Monte Carlo variability as in Table 1. Nevertheless, across all models and both platforms, the TLCQM-based approaches consistently outperforms the target-only ML models and other competing transfer learning methods, even after accounting for this additional randomness. These results demonstrate the robustness and practical relevance of the proposed framework in a business-critical application.

5 Discussion

In summary, this work presents a transfer learning framework that combines source-domain conditional generators through a shared calibration estimated at the observed target covariates. Our theoretical and empirical results show that the framework can generate useful synthetic regression data for data-scarce target domains, while also revealing the risk of negative transfer when the target conditional distribution is poorly approximated by the calibrated family induced by the source generators. Although our analysis focuses on regression with univariate continuous responses, the conditional quantile calibration principle can be extended to other settings. For multivariate continuous responses, the quantile matching estimator in (7) can be generalized via a weighted sum of CRPS objectives or a multivariate proper score, such as the energy score [Gneiting and Raftery, 2007]. For binary or categorical responses in classification tasks, CRPS coincides with the Brier score. More broadly, the distributional calibration step in our proposed framework could be implemented using alternative metrics or operators, such as the sliced Wasserstein metric [Bonneel et al., 2015], Sinkhorn distance [Cuturi, 2013], or maximum mean discrepancy [Cui et al., 2020]. A systematic investigation of these extensions is left for future work.

Acknowledgments

The authors thank Xu Chen, Rizen Yamauchi, and other members of the Central Applied Science as well as Research Platform teams at Meta for their support and valuable feedback throughout this research.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted.

References

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the rademacher complexity of linear hypothesis sets. *arXiv preprint arXiv:2007.11045*, 2020.

Ruiqi Bai, Yijiao Zhang, Hanbo Yang, and Zhongyi Zhu. Transfer learning for high-dimensional quantile regression with distribution shift. *arXiv preprint arXiv:2411.19933*, 2024.

Ludwig Baringhaus and Carsten Franz. On a new multivariate two-sample test. *Journal of Multivariate Analysis*, 88(1):190–206, 2004.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of machine learning research*, 3(Nov):463–482, 2002.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

T Tony Cai and Hongming Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*, 2024.

T Tony Cai and Hongji Wei. Transfer learning for nonparametric classification. *The Annals of Statistics*, 49(1):100–128, 2021.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi, editors, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM, 2016.

Xiaohong Chen and Timothy Christensen. Optimal uniform convergence rates for sieve nonparametric instrumental variables regression. *arXiv preprint arXiv:1311.0412*, 2013.

Xinyang Chen, Sinan Wang, Jianmin Wang, and Mingsheng Long. Representation subspace distance for domain adaptation regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1749–1759. PMLR, Jul 2021.

Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45(1):223 – 256, 2017.

Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33:17164–17175, 2020.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Yves Dominicy and David Veredas. The method of simulated quantiles. *Journal of Econometrics*, 172(2):235–247, 2013.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International Conference on Machine Learning*, pages 2839–2848. PMLR, 2016.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, Bernhard Schölkopf, et al. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4): 5, 2009.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Baihua He, Huihang Liu, Xinyu Zhang, and Jian Huang. Representation transfer learning for semiparametric regression. *arXiv preprint arXiv:2406.13197*, 2024.

Hans Hersbach. Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570, 2000.

Masaaki Imaizumi. Sup-norm convergence of deep neural network estimator for nonparametric regression by adversarial training. *arXiv preprint arXiv:2307.04042*, 2023.

Jun Jin, Jun Yan, Robert H Aseltine, and Kun Chen. Transfer learning with large-scale quantile regression. *Technometrics*, 66(3):381–393, 2024.

Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *Advances in neural information processing systems*, 21, 2008.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. Apartment for rent classified. UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5X623.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Vladimir Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: Ecole D'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011.

Michael R Kosorok. Two-sample quantile tests under general conditions. *Biometrika*, 86(4):909–921, 1999.

Jean-Samuel Leboeuf, Frédéric LeBlanc, and Mario Marchand. Decision trees as partitioning machines to characterize their generalization properties. *Advances in neural information processing systems*, 33:18135–18145, 2020.

Sai Li, T Tony Cai, and Hongzhe Li. Transfer learning for high-dimensional linear regression: Prediction, estimation and minimax optimality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(1):149–173, 2022.

Haotian Lin and Matthew Reimherr. Smoothness adaptive hypothesis transfer learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 30286–30316. PMLR, 21–27 Jul 2024.

Haotian Lin and Matthew Reimherr. Model-robust and adaptive-optimal transfer learning for tackling concept shifts in nonparametric regression. *arXiv preprint arXiv:2501.10870*, 2025.

Xu Liu, Yingguang Li, Qinglu Meng, and Gengxiang Chen. Deep transfer learning for conditional shift in regression. *Knowledge-Based Systems*, 227:107216, 2021.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105. PMLR, 2015.

Colin L Mallows. A note on asymptotic joint normality. *The Annals of Mathematical Statistics*, 43 (2):508–515, 1972.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation with multiple sources. *Advances in neural information processing systems*, 21, 2008.

Pascal Massart. Some applications of concentration inequalities to statistics. *Annales de la Faculté des sciences de Toulouse: Mathématiques*, 9(2):245–303, 2000.

James E Matheson and Robert L Winkler. Scoring rules for continuous probability distributions. *Management Science*, 22(10):1087–1096, 1976.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, second edition, 2018.

Whitney K Newey and James R Robins. Cross-fitting and fast remainder rates for semiparametric estimation. *arXiv preprint arXiv:1801.09138*, 2018.

Tuan Duong Nguyen, Marthinus Christoffel, and Masashi Sugiyama. Continuous target shift adaptation in supervised learning. In *Asian Conference on Machine Learning*, pages 285–300. PMLR, 2016.

XuanLong Nguyen, Martin J Wainwright, and Michael Jordan. Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, 2007.

Rajbir-Singh Nirwan and Nils Bertschinger. Bayesian quantile matching estimation. *arXiv preprint arXiv:2008.06423*, 2020.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE transactions on neural networks*, 22(2):199–210, 2010.

Sheng Qiao, Yong He, and Wenxin Zhou. Transfer learning for high-dimensional quantile regression with statistical guarantee. *Transactions on Machine Learning Research*, 2024.

Sashank Reddi, Barnabas Poczos, and Alex Smola. Doubly robust covariate shift correction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 29, 2015.

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation*, 14(1):21–41, 2002.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Jessica Schrouff, Natalie Harris, Sanmi Koyejo, Ibrahim M Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. *Advances in Neural Information Processing Systems*, 35:19304–19318, 2022.

Nikolaos Sgouropoulos, Qiwei Yao, and Claudia Yastremiz. Matching a distribution by matching quantiles estimation. *Journal of the American Statistical Association*, 110(510):742–759, 2015.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3): 653–677, 2025.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.

Charles J Stone. Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics*, pages 1348–1360, 1980.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Zahra Taghiyarrenani, Sławomir Nowaczyk, Sepideh Pashami, and Mohamed-Rafik Bouguelia. Multi-domain adaptation for regression under conditional distribution shift. *Expert systems with applications*, 224:119907, 2023.

Pham Dinh Tao and Le Thi Hoai An. A dc optimization algorithm for solving the trust-region subproblem. *SIAM Journal on Optimization*, 8(2):476–505, 1998.

Pham Dinh Tao and LT Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta Mathematica Vietnamica*, 22(1):289–355, 1997.

Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.

Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global Scientific Publishing, 2010.

Lan V Truong. On rademacher complexity-based generalization bounds for deep learning. *arXiv preprint arXiv:2208.04284*, 2022.

Katherine Tsai, Stephen R Pfohl, Olawale Salaudeen, Nicole Chiou, Matt Kusner, Alexander D'Amour, Sanmi Koyejo, and Arthur Gretton. Proxy methods for domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3961–3969. PMLR, 2024.

Rosanna Turrisi, Rémi Flamary, Alain Rakotomamonjy, and Massimiliano Pontil. Multi-source domain adaptation via weighted joint distributions optimal transport. In *Uncertainty in artificial intelligence*, pages 1970–1980. PMLR, 2022.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2008.

Chao Wang, Caixing Wang, Xin He, and Xingdong Feng. Transfer learning for kernel-based regression. *arXiv preprint arXiv:2310.13966*, 2023.

Kaizheng Wang. Pseudo-labeling for kernel ridge regression under covariate shift. *arXiv preprint arXiv:2302.10160*, 2023.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11293–11302, 2019.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69(2):297–324, 2018.

Steven Wilkins-Reeves, Xu Chen, Qi Ma, Christine Agarwal, and Aude Hofleitner. Multiply robust estimation for local distribution shifts with multiple domains. In *International Conference on Machine Learning*, pages 52972–52993. PMLR, 2024.

Yujie Wu, Giovanni Parmigiani, and Boyu Ren. Multi-source domain adaptation for regression. *arXiv preprint arXiv:2312.05460*, 2023.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pages 819–827. Pmlr, 2013.

Kun Zhang, Mingming Gong, and Bernhard Schölkopf. Multi-source domain adaptation: A causal view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), 2015.

Wen Zhang, Lingfei Deng, Lei Zhang, and Dongrui Wu. A survey on negative transfer. *IEEE/CAA Journal of Automatica Sinica*, 10(2):305–329, 2022.

Yijiao Zhang and Zhongyi Zhu. Transfer learning for high-dimensional quantile regression via convolution smoothing. *arXiv preprint arXiv:2212.00428*, 2022.

Doudou Zhou, Mengyan Li, Tianxi Cai, and Molei Liu. Model-assisted and knowledge-guided transfer regression for the underrepresented population. *arXiv preprint arXiv:2410.06484*, 2024.

Doudou Zhou, Molei Liu, Mengyan Li, and Tianxi Cai. Doubly robust augmented model accuracy transfer inference with high dimensional features. *Journal of the American Statistical Association*, 120(549):524–534, 2025.

The first row in Table 2 is also known as the finite class lemma or Massart's lemma [Massart, 2000]. By Sauer's Lemma, if the VC dimension $\mathrm{VC}(\mathcal{F})$ of a bounded real-valued function class $\mathcal{F}$ is less than n, then $\mathrm{Rad}_n(\mathcal{F}) = O\left(\sqrt{\frac{\mathrm{VC}(\mathcal{F})\log(2n)}{n}}\right)$.

Function Class	Upper Bound	References
Finite hypothesis class $\mathcal{F}$ with $\sup_{x\in\mathcal{X}} \lvert f(x)\rvert \le B_f$ for any $f \in \mathcal{F}$.	$O\left(B_f\sqrt{\frac{\log(2\lvert\mathcal{F}\rvert)}{n}}\right)$	Theorem 3.7 in Mohri et al. [2018]; see also Massart [2000].
Linear predictor class $\mathcal{F} = \{x \mapsto \Upsilon^T x : \lVert\Upsilon\rVert_2 \le B_\gamma, \lVert x\rVert_2 < B_x\}$.	$O\left(\frac{B_\gamma B_x}{\sqrt{n}}\right)$	Theorem 3 in Kakade et al. [2008]; see also Theorem 1 in Awasthi et al. [2020].
Sparse linear predictor $\mathcal{F} = \{x \mapsto \Upsilon^T x : x \in \mathbb{R}^d, \lVert\Upsilon\rVert_0 \le s,\ \lVert\Upsilon\rVert_2 \le B_\gamma,\ \lVert x\rVert_2 < B_x\}$.	$O\left(B_\gamma B_x\sqrt{\frac{s\log d}{n}}\right)$	Section 3.1 in Kakade et al. [2008]; see also Lemma 7.1 in Koltchinskii [2011].
Reproducing kernel Hilbert space $\mathcal{F} = \{f \in \mathcal{H} : \lVert f\rVert_{\mathcal{H}} \le B_f\}$ with $\sup_x \sqrt{k(x,x)} \le \kappa$	$O\left(\frac{B_f\kappa}{\sqrt{n}}\right)$	Lemma 22 in Bartlett and Mendelson [2002].
Binary decision tree class with L nodes and d real-valued features	$O\left(\sqrt{\frac{L\log(2Ld)\log(2n)}{n}}\right)$	Corollary 10 in Leboeuf et al. [2020].
Neural network with the ReLU activation function, W weights, L layers, and outputs clipped to $[-B_f, B_f]$.	$O\left(\sqrt{\frac{WL\log(2W)\log(2n)}{n}}\right)$	Theorem 6 in Bartlett et al. [2019]; see also Theorem 5 in Yin et al. [2019], Theorem 14 in Truong [2022], and Theorem 3.3 in Bartlett et al. [2017].

Table 2: Upper bounds on the Rademacher complexity $\mathrm{Rad}_n(\mathcal{F})$ for common function classes

A Other Related Work

Transfer learning: Within the broad literature on transfer learning [Pan and Yang, 2009], our work falls under supervised or transductive transfer learning for regression, also known as multi-domain adaptation [Mansour et al., 2008]. We therefore do not review related but distinct lines of work on transfer learning for classification [Cai and Wei, 2021] or unsupervised domain adaptation [Ganin et al., 2016, Long et al., 2015, Chen et al., 2021]. Recent theoretical advances in supervised transfer learning for regression have focused on high-dimensional (generalized linear) models [Li et al., 2022, Tian and Feng, 2023, Zhou et al., 2024, 2025], kernel ridge regression [Wang et al., 2023, Wang, 2023, Lin and Reimherr, 2024, 2025], and more general nonparametric settings [Cai and Pu, 2024]. The generalized target shift scenario (1) has also been studied through different angles, including location-scale conditional shift models [Zhang et al., 2013], conditional embedding operators from the maximum mean discrepancy in deep learning [Liu et al., 2021], outcome coarsening combined with representation learning [Wu et al., 2023], pairwise similarity-preserving feature extraction to correct distributional discrepancies in $X|Y$ [Taghiyarrenani et al., 2023], and semiparametric regression models with representation learning [He et al., 2024]. Our focus instead is conditional distribution calibration for data augmentation under the general shift setting (1).

Quantile matching: Quantile matching [Sgouropoulos et al., 2015] has appeared in various contexts, including two-sample testing [Kosorok, 1999], parameter estimation [Dominicy and Veredas, 2013], model diagnostics via quantile-quantile plots [Stuart, 2010], and Bayesian methods [Nirwan and Bertschinger, 2020]. More broadly, quantile matching can be interpreted as a special case of solving an optimal transport problem between univariate marginal distributions [Mallows, 1972, Villani, 2008].

Finally, we emphasize that our work is distinct from transfer learning for quantile regression [Zhang and Zhu, 2022, Qiao et al., 2024, Bai et al., 2024, Jin et al., 2024], which aims to improve estimation and inference for conditional quantiles in the target domain by leveraging source data. In contrast, our framework focuses on data augmentation through distributional alignment and can be used as a general preprocessing step to enhance a wide range of downstream learning tasks, including but not limited to quantile regression.

B Additional Implementation Details

In this section, we document the practical considerations involved in implementing our proposed TLCQM framework and other experimental setups.

B.1 Computing Environment

For parallelizing our Monte Carlo experiments, our code was executed on a computing clusters with 82 compute nodes of CPU cores in total. The required memory for each job is clearly stated in our code.

B.2 Setups for Engression and Conditional Quantile Matching

For the engression method described in Section 2.1, we employ a neural network model with two hidden layers, each consisting of 100 neurons. The latent noise dimension is set to 5, the learning rate to 0.001, and the model is trained for 1000 epochs unless otherwise specified.

For the conditional calibration estimator in Section 2.2, we generate $M = 3000$ responses from each fitted source conditional generative model at every observed target covariate. Unless otherwise stated, we minimize (7) over

$$\Theta = \left\{\boldsymbol{\beta} \in \mathbb{R}^{K+1} : \|\boldsymbol{\beta}\|_1 \leq 10\right\}.$$

For a fixed $\boldsymbol{\beta}$, we define $\widehat{Z}_{ij}(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \widehat{\boldsymbol{V}}_{ij}$, and let $\widehat{Z}_{i(1)}(\boldsymbol{\beta}) \leq \cdots \leq \widehat{Z}_{i(M)}(\boldsymbol{\beta})$ denote the corresponding order statistics for target observation i. The pairwise term in (7) can then be evaluated as

$$\frac{1}{2M(M-1)} \sum_{j \neq j'} \left|\widehat{Z}_{ij}(\boldsymbol{\beta}) - \widehat{Z}_{ij'}(\boldsymbol{\beta})\right| = \frac{1}{M(M-1)} \sum_{r=1}^{M} (2r - M - 1)\widehat{Z}_{i(r)}(\boldsymbol{\beta}).$$

Consequently, evaluating the empirical CRPS and a corresponding subgradient costs $O\left(n_0 M(K + \log M)\right)$. Crucially, the M responses are sorted separately for each i so that responses generated at different target covariates are neither compared nor pooled.

We initialize the optimization by regressing the observed target responses on the Monte Carlo conditional means of the fitted source models and projecting the resulting coefficient vector onto Θ as:

$$\widetilde{\boldsymbol{\beta}} = \Pi_\Theta \left[\underset{\boldsymbol{\beta} \in \mathbb{R}^{K+1}}{\arg\min} \sum_{i=1}^{n_0} \left\{ Y_i^{(0)} - \beta_0 - \sum_{k=1}^{K} \beta_k \left(\frac{1}{M} \sum_{j=1}^{M} \widehat{Y}_{ij}^{(k)} \right) \right\}^2 \right],$$

where Π_Θ is the projection operator onto the set Θ. Because (7) is generally nonconvex, we supplement this least-squares initialization with 9 random feasible initializations and retain the terminal solution having the smallest finite empirical CRPS. For each initialization, we apply the proximal difference-of-convex update (8) with $\gamma = 10^{-4}$. At each outer iteration, the convex surrogate is solved numerically using sequential least-squares programming (SLSQP) with at most 1000 inner iterations. A candidate solution is accepted only when it also does not increase the original empirical CRPS. The algorithm runs for at most 1000 outer iterations and terminates when either

$$\left\|\boldsymbol{\beta}^{(r+1)} - \boldsymbol{\beta}^{(r)}\right\|_2 \leq 10^{-8} \left(1 + \left\|\boldsymbol{\beta}^{(r)}\right\|_2\right)$$

or

$$\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}^{(r)}) - \widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}^{(r+1)}) \leq 10^{-8} \left[1 + \left|\widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}^{(r+1)})\right|\right].$$

For squared-error prediction, we use the calibrated conditional mean in (9), rather than an individual synthetic response, as the augmented label. The separate Monte Carlo size M' used to approximate each source conditional mean is set to $M' = 200$.

B.3 Method for Density Ratio Estimation

In Step 5 of Algorithm 1, we estimate the density ratio $w_k(x) = \frac{dP_X^{(0)}(x)}{dP_X^{(k)}(x)}$ for $k = 1, ..., K$ using kernel mean matching [Gretton et al., 2009, 2012]. Let n_0 and $n_k, k = 1, ..., K$ denote the sample

sizes of the target and source domains, respectively. For each source domain k, given a positive definite kernel function $G : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define the matrix $G^{(k)} \in \mathbb{R}^{n_k \times n_k}$ by $G_{ij}^{(k)} := G(X_i^{(k)}, X_j^{(k)})$ and the vector $\kappa^{(k)} \in \mathbb{R}^{n_k}$ by $\kappa_i^{(k)} = \frac{n_k}{n_0} \sum_{j=1}^{n_0} G(X_i^{(k)}, X_j^{(0)})$. Kernel mean matching estimates the sample density-ratio vector $\zeta^{(k)} = \left(w_k(X_1^{(k)}), ..., w_k(X_{n_k}^{(k)})\right)^T \in \mathbb{R}^{n_k}$ by solving

$$\begin{aligned} \min_{\zeta^{(k)} \in \mathbb{R}^{n_k}} \quad & \frac{1}{n_k^2} \zeta^{(k)T} G^{(k)} \zeta^{(k)} - \frac{2}{n_k^2} \zeta^{(k)T} \kappa^{(k)} \\ \text{subject to} \quad & 0 \le \zeta_i^{(k)} \le B_\zeta, \qquad \left| \sum_{i=1}^{n_k} \zeta_i^{(k)} - n_k \right| \le n_k \xi, \end{aligned}$$

where ξ is often chosen to be $O\left(\frac{B_\zeta}{\sqrt{n_k}}\right)$.

B.4 Hyperparameter Tuning for Standard Machine Learning Models and Comparative Methods

When we apply the standard machine learning methods to datasets under various scenarios, their hyperparameters are tuned by five-fold cross-validation across a combination of the following candidate choices:

- **XGBoost:** `learning_rate`: $[0.001, 0.01, 0.1]$, `n_estimators`: $[10, 50, 100]$, `max_depth`: $[3, 5]$, `subsample`: $[0.8, 1.0]$, `colsample_bytree`: $[0.8, 1.0]$.
- **Kernel Ridge Regression (KRR):** `alpha` (penalty parameter): $\{3^r : r = -2, ..., 5\} \times \frac{0.1}{n}$, where n is the sample size of the entire training set.
- **Neural network:** `hidden_layer_sizes`: $[(10,), (50,), (100,)]$ and `alpha` (L_2 regularization strength): $[0.0001, 0.001, 0.01]$ with the activation function as "ReLU", the optimizer as Adam [Kingma and Ba, 2014], and the number of epochs as 1000.

For transfer learning with kernel ridge regression (TKRR; Wang et al. 2023), we likewise use five-fold cross-validation for selecting the optimal penalty parameter `alpha` within the candidate set $\left[3^{-3}, 3^{-1}, ..., 3^7\right] \times \frac{0.1}{n}$, where n is the sample size of the entire training set. For deep transfer learning for conditional shift in regression (CDAR; Liu et al. 2021), we follow the same setup in their paper, except that we replace the convolution architecture with a fully-connected neural network with $d \times 64 \times 16 \times 8$ hidden neurons. Finally, for multi-domain adaptation for regression under conditional shift (DARC; Taghiyarrenani et al. 2023), we set the embedding or latent feature dimension across domains as 8, whose feature extraction model is a neural network with $(d + 8) \times 100 \times 100$ hidden neurons.

C Additional Simulation Results

This section provides supplementary simulation results for other baseline training strategies of the machine learning models, ablation studies for our proposed TLCQM framework, and diagnostic tests for negative transfer under other simulation setups.

C.1 Other Baseline Training Strategies

We further compare the performance of XGBoost, KRR, and NN under several training strategies within the simulation framework described in Section 4.1. Specifically, we compare models trained using (i) source-only data, (ii) naive pooling of source and target data, (iii) target-only data augmentation via the "mixup" technique [Zhang et al., 2018], and (iv) data augmentation via TLCQM. The "mixup" strategy generates additional target-domain observations by forming convex combinations of randomly selected pairs of target samples. Given two observations $(X_i^{(0)}, Y_i^{(0)})$ and $(X_j^{(0)}, Y_j^{(0)})$ randomly drawn from the original target-only training data, "mixup" creates the synthetic observation

$$\widetilde{X}^{(0)} = \Lambda X_i^{(0)} + (1 - \Lambda) X_j^{(0)}, \quad \widetilde{Y}^{(0)} = \Lambda Y_i^{(0)} + (1 - \Lambda) Y_j^{(0)},$$

Ratio $\frac{n_k}{n_0}$	Source 1 only	Source 2 only	Target only	Naive pooling	"Mixup" (best $\widetilde{\alpha}$)	TLCQM
XGBoost						
1	4.317	4.323	**0.307**	0.619	0.354	0.315
2	4.320	4.270	**0.307**	1.221	0.354	0.314
5	4.357	4.158	**0.307**	2.054	0.355	0.308
10	4.485	4.226	0.307	2.421	0.354	**0.304**
15	4.666	4.397	0.307	2.494	0.354	**0.303**
20	4.809	4.485	0.307	2.516	0.354	**0.302**
30	4.967	4.629	0.307	2.545	0.355	**0.302**
Kernel Ridge Regression (KRR)						
1	4.128	3.741	0.328	0.550	1.314	**0.326**
2	5.174	4.411	0.328	0.679	1.314	**0.322**
5	6.308	7.469	0.328	0.993	1.313	**0.315**
10	6.539	10.146	0.328	1.406	1.313	**0.309**
15	6.459	11.180	0.328	1.724	1.314	**0.305**
20	6.300	11.833	0.328	1.983	1.312	**0.303**
30	5.984	12.657	0.328	2.407	1.311	**0.300**
Neural Network (NN)						
1	5.212	6.102	0.375	0.628	0.417	**0.322**
2	5.009	5.787	0.375	0.721	0.418	**0.316**
5	4.907	6.366	0.375	1.016	0.418	**0.309**
10	4.901	7.354	0.375	1.385	0.418	**0.304**
15	4.908	7.730	0.375	1.710	0.417	**0.302**
20	4.847	7.839	0.375	2.013	0.417	**0.302**
30	4.738	7.789	0.375	2.498	0.417	**0.301**

Table 3: Prediction performance of machine learning models under different training strategies and source-to-target ratios. For each model and data scenario, the "mixup" column reports the best-performing choice of $\widetilde{\alpha}$ in terms of MSE.

where $\Lambda \sim \mathrm{Beta}(\widetilde{\alpha}, \widetilde{\alpha})$ with tuning parameter $\widetilde{\alpha} \in (0, 1)$. In our experiments, we set the number of "mixup"-generated samples to $5n_0$ with n_0 being the target-only sample size. For each machine learning model and data scenario, the tuning parameter $\widetilde{\alpha}$ is selected from the candidate set $\{0.2, 0.4, 0.5, 0.6, 0.8\}$ to minimize the mean squared error (MSE).

C.2 Ablation Studies

The experiments in Section C.1 already provide preliminary ablation evidence on the importance of proper transfer learning strategies, where training machine learning models on source-only data, target-only data, naively pooling all available data, or simple target-only data augmentation often leads to inferior performances. Here, we conduct more systematic ablation experiments under the simulation setup in Section 4.1 to clarify the contributions of the key components in our TLCQM framework. Specifically, we compare four different variants: (i) engression only (no calibration), (ii) conditional mean matching (MM), (iii) optimal transport (OT) calibration based on empirical cumulative distribution functions, and (iv) the proposed TLCQM framework. Notice that the calibration procedures in variants (ii-iv) are applied on top of the generated responses from the same fitted engression model.

To evaluate the performances of these four variants, we compute the MSEs of XGBoost, KRR, and NN trained on the generated or calibrated responses, as well as the Wasserstein distance between the generated or calibrated responses and the true responses in the target domain. Notably, smaller marginal Wasserstein distances imply better distributional alignment between the generated responses and the true target responses, which often benefits downstream machine learning tasks not only for point prediction but also for quantile regression and distributional learning.

Source-to-Target Ratio $\frac{n_k}{n_0}$	1	2	5	10	15	20	30
Mean Squared Error (XGBoost)							
Engression Only	0.880	2.420	2.965	3.209	3.319	3.374	3.557
Conditional MM	**0.307**	**0.309**	**0.307**	0.305	0.303	0.303	0.302
OT Calibration	0.357	0.379	0.421	0.470	0.507	0.537	0.587
TLCQM	0.315	0.314	**0.307**	**0.304**	**0.303**	**0.302**	**0.302**
Mean Squared Error (KRR)							
Engression Only	0.757	0.959	1.369	1.816	2.156	2.424	2.821
Conditional MM	**0.319**	**0.318**	**0.314**	0.310	0.306	0.304	0.301
OT Calibration	0.371	0.394	0.450	0.514	0.557	0.586	0.625
TLCQM	0.326	0.322	0.315	**0.309**	**0.305**	**0.303**	**0.300**
Mean Squared Error (NN)							
Engression Only	0.926	1.610	2.490	2.940	3.175	3.329	3.576
Conditional MM	**0.314**	**0.312**	**0.308**	0.305	0.303	0.303	0.302
OT Calibration	0.386	0.411	0.456	0.535	0.597	0.638	0.687
TLCQM	0.322	0.316	0.309	**0.304**	**0.302**	**0.302**	**0.301**
Wasserstein Distance							
Engression Only	1.999	2.001	2.006	2.015	2.022	2.028	2.032
Conditional MM	0.346	0.342	0.327	0.316	0.309	0.306	0.298
OT Calibration	0.566	0.539	0.463	0.387	0.341	0.323	0.294
TLCQM	**0.0736**	**0.0530**	**0.0486**	**0.0479**	**0.0478**	**0.0479**	**0.0481**

Table 4: Results for ablation studies comparing engression only, conditional mean matching, OT calibration, and the proposed TLCQM framework across different source-to-target ratios.

As shown by the ablation results in Table 4, the engression model alone is insufficient for effective transfer learning, indicating that the performance gains of our TLCQM framework are not merely due to a strong generative model. In contrast, all calibrated variants perform substantially better, confirming that distributional calibration is indispensable. Furthermore, simple mean matching and TLCQM achieve comparable MSEs for downstream machine learning models in this simulation. Nevertheless, TLCQM substantially outperforms both simple mean matching and OT calibration in terms of distributional alignment, as reflected by its much smaller reported Wasserstein distances to the true target responses than the other two variants. These ablation studies further highlight the integral design of our TLCQM framework, where both generative models and the quantile matching estimator play essential roles.

C.3 Diagnostic Studies

We further conduct two diagnostic simulations to distinguish marginal from conditional distributional alignment and to examine when TLCQM data augmentation may lead to negative transfer. In both settings below, $X = (X_1, ..., X_5)^T \sim \mathcal{N}(\mathbf{0}_5, \boldsymbol{I}_5)$ in every domain, and the response noise is independently drawn from $\epsilon \sim \mathcal{N}(0, 0.3^2)$.

- **Reversed conditional relation:** For the two candidate source domains, $Y^{(1)} = -X_1 + \epsilon$ and $Y^{(2)} = X_2 + \epsilon$. For the target domain, $Y^{(0)} = X_1 + \epsilon$. Thus, the target conditional distribution belongs to the calibrated family in (7) with $\boldsymbol{\beta} = (0, -1, 0)^T$. Since X_1 is symmetric around 0, the target and first source domains have identical marginal response distributions, but their conditional distributions are reversed.
- **Outside affine span:** For the two candidate source domains, $Y^{(1)} = \sin(\theta^T X) + \epsilon$ and $Y^{(2)} = \cos(\theta^T X) + \epsilon$, where $\theta = \left(1, \frac{1}{2}, ..., \frac{1}{5}\right)^T$. For the target domain, $Y^{(0)} = \sin(2\theta^T X) + \epsilon$. Hence, the target conditional mean lies outside the affine span generated by the two source conditional means, as specified in (7) and further described in Remark 2.1.

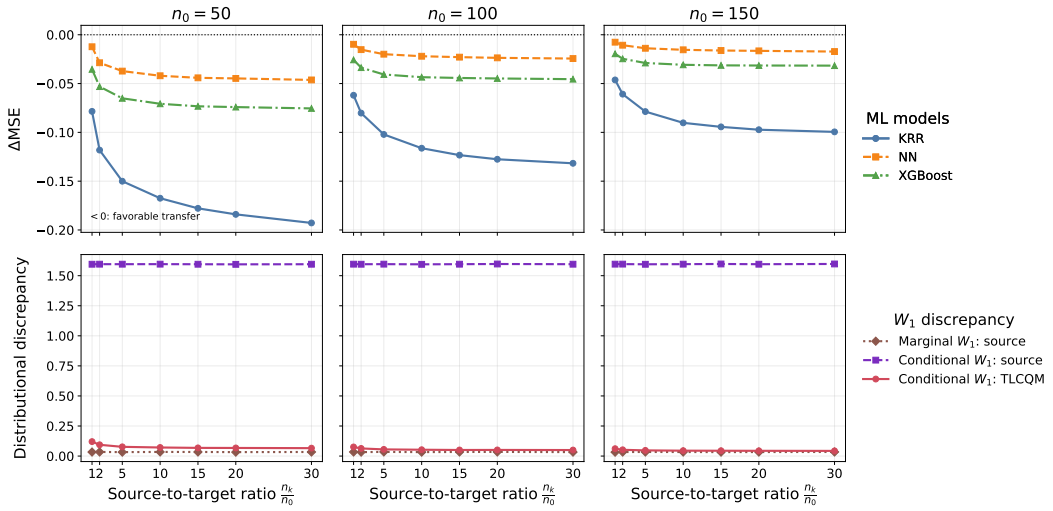


Figure 3: Simulation results under the "**reversed conditional relation**" scenario. The top row reports $\Delta\text{MSE} = \text{MSE}_{\text{TLCQM}} - \text{MSE}_{\text{Target}-\text{only}}$; negative values indicate favorable transfer. The bottom row reports marginal and conditional Wasserstein-1 distances from the first source and TLCQM-calibrated distributions to the target one.

We vary the target-domain sample size n_0 from 50 to 150 and adjust the source-to-target ratio $\frac{n_k}{n_0}$ between source and target sample sizes from 1 to 30 for $k = 1, 2$. For each downstream ML model (XGBoost, KRR, NN), we report the MSE differences

$$\Delta\text{MSE} = \text{MSE}_{\text{TLCQM}} - \text{MSE}_{\text{Target}-\text{only}},$$

so negative values indicate favorable transfer. We also compute marginal and conditional Wasserstein-1 distances between the target distribution and either the first source distribution or the TLCQM-calibrated distribution. These Wasserstein-1 metrics allow us to assess the distributional discrepancies and intrinsic causes of negative transfer. At each of 3000 independent target covariates, we generate 500 responses from the known source and target conditional distributions. The marginal Wasserstein distance pools responses across all test covariates, whereas the conditional Wasserstein distance is computed separately at each covariate and then averaged.

The results are shown in Figure 3 and Figure 4. In the "**reversed conditional relation**" setting, the target and first source distributions have small marginal Wasserstein-1 distances but large conditional Wasserstein-1 distances across the considered values of n_0 and n_k. This reveals that marginal alignment can conceal an incorrect covariate-response relationship. In contrast, TLCQM recovers the conditional reversal and reduces the conditional Wasserstein-1 distances between TLCQM-calibrated and target responses to nearly zero. As a result, the downstream ML models trained on TLCQM-augmented data achieve lower MSEs than their target-only counterparts, indicating favorable transfer. In the "**outside affine span**" setting, the marginal Wasserstein-1 distances are again small, but the conditional discrepancies of both source and TLCQM-calibrated distributions to the target one remain large. Since a shared linear calibration (3) cannot eliminate this approximation bias, the downstream ML models trained on the TLCQM-augmented data often exhibit negative transfer. These results illustrate the bias-variance tradeoff in Theorem 3.4: conditional quantile calibration (7) can correct misleading marginal distributional alignment when the target distribution belongs to the calibrated family as in (3), but persistent transfer bias can dominate the generalization benefit of additional source data when the target conditional distribution lies outside this linear span.

D Experiments on Real Public Data

To showcase the effectiveness of our proposed TLCQM framework on real-world data, we next evaluate it on a public dataset `Apartment` from the UCI machine learning repository [Kelly et al., 2019]. The response variable Y is defined as the logarithm of the apartment rental price. The covariate vector X comprises 10 features, including the numbers of bathrooms and bedrooms, the apartment

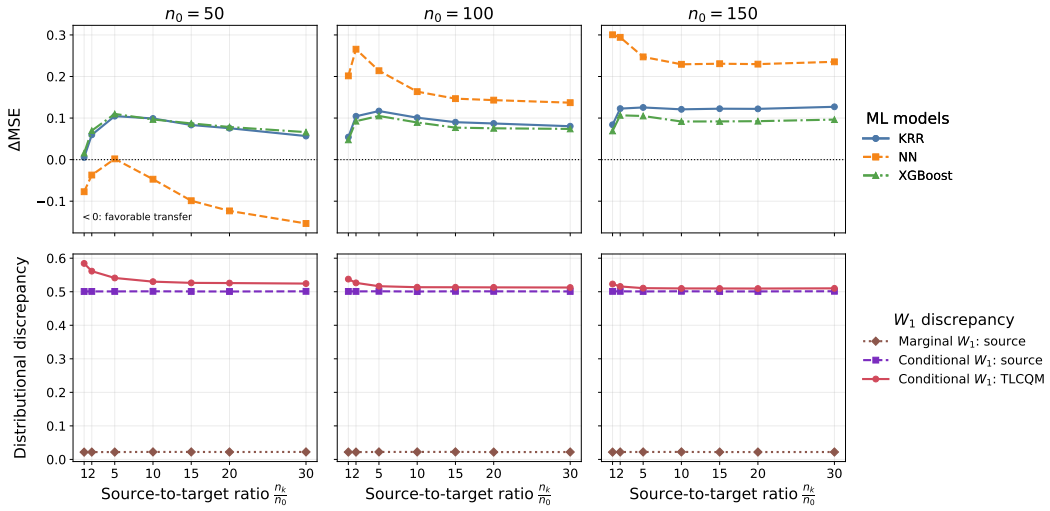


Figure 4: Simulation results under the "**outside affine span**" scenario. The top row reports $\Delta\text{MSE} = \text{MSE}_{\text{TLCQM}} - \text{MSE}_{\text{Target}-\text{only}}$; negative values indicate favorable transfer. The bottom row reports marginal and conditional Wasserstein-1 distances from the first source and TLCQM-calibrated distributions to the target one.

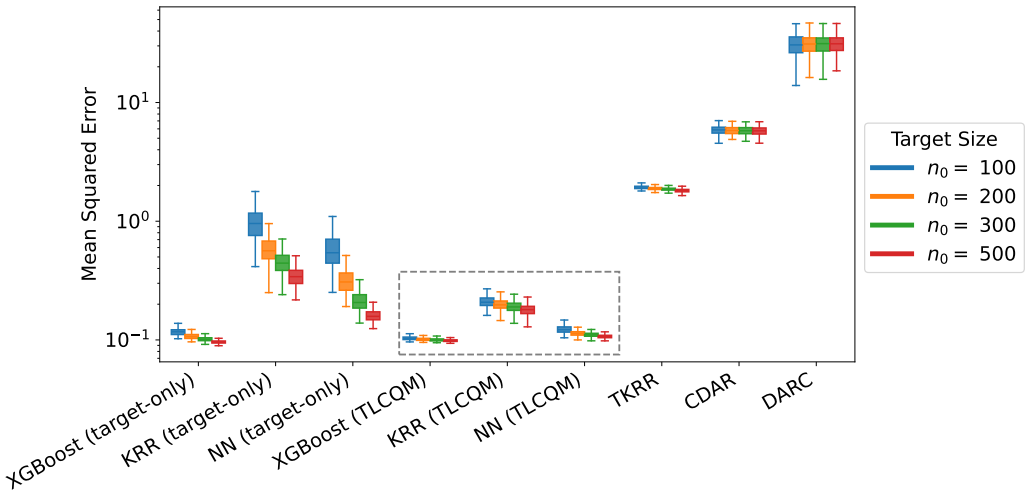


Figure 5: Performances of different machine learning models applied to target-only and TLCQM scenarios as well as other competing transfer learning methods on the "Apartment" data.

size in square feet, and 7 binary indicators capturing the presence of photos, parking, storage, a gym, a pool, and whether cats and dogs are permitted. The source domains consist of observations from three randomly selected states: `IL` ($n_1 = 1036$), `OH` ($n_2 = 4905$), and `WA` ($n_3 = 2612$). The target domain is the state `FL`, which contains 5775 observations in total. To construct the training data for the target domain, we randomly subsample $n_0 \in \{100, 200, 300, 500\}$ observations from `FL` and reserve the remaining observations as the test set.

The results, summarized in Figure 5 and Table 5 based on 500 Monte Carlo replications, show that machine learning models trained under our proposed TLCQM framework consistently have smaller MSEs than their target-only counterparts as well as other competing transfer learning methods. The performance gains are particularly pronounced for NN, exceeding those observed for KRR and XGBoost. This pattern is consistent with the mechanism of our TLCQM framework, which improves predictive performance primarily by enlarging the effective training sample size through

n_0	100	200	300	500
XGBoost	11.99%	5.49%	1.70%	-2.70%
KRR	78.49%	66.00%	57.89%	47.81%
NN	80.01%	65.86%	49.55%	34.10%

Table 5: Percentage reduction in MSE achieved by machine learning models trained on TLCQM-augmented data relative to their target-only counterparts on the "Apartment" data.

the generation of well-calibrated synthetic data, thereby reducing estimation variance. Consequently, flexible learners whose performance is strongly sample-size dependent, such as NN and KRR, benefit more from the TLCQM framework. In contrast, tree-based models often exhibit stable performance in small-sample tabular settings [Shwartz-Ziv and Armon, 2022], leaving less room for variance reduction to translate into substantial gains.

E Proofs for the Conditional Quantile Calibration Results

E.1 Conditional CRPS Identity

For any CDF G with a finite first moment, we recall from (4) that

$$\mathrm{CRPS}(G, y) = \int_{\mathbb{R}} \left[G(t) - \mathbb{1}\{y \le t\}\right]^2 dt.$$

Conditioning on $X = x$, by Tonelli's theorem and using $\mathbb{P}(Y^{(0)} \le t \mid X = x) = F^{(0)}(t \mid x)$, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\mathrm{CRPS}(G(\cdot|X), Y^{(0)})\middle|X = x\right] &= \int_{\mathbb{R}} \mathbb{E}\left[\left(G(t|x) - \mathbb{1}\{Y^{(0)} \le t\}\right)^2 \middle| X = x\right] dt \\
&= \int_{\mathbb{R}} \left[G(t|x)^2 - 2F^{(0)}(t|x) \cdot G(t|x) + F^{(0)}(t|x)\right] dt \\
&= \int_{\mathbb{R}} \left[G(t|x) - F^{(0)}(t|x)\right]^2 dt + \int_{\mathbb{R}} F^{(0)}(t|x)\left[1 - F^{(0)}(t|x)\right] dt.
\end{aligned}$$

Additionally, (6) can be proved by integrating over X and noting that

$$\mathbb{E}\left[\mathrm{CRPS}(F^{(0)}(\cdot|X), Y^{(0)})|X = x\right] = \int_{\mathbb{R}} F^{(0)}(t|x)\left[1 - F^{(0)}(t|x)\right] dt.$$

E.2 Proof of Theorem 3.4

Theorem 3.4. Suppose that Assumptions 3.1 and 3.2 hold and that $\left(\widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}}\right)$ is fitted on auxiliary data independent of the observations used in (11). Assume further that $\ell(y, y') = (y - y')^2$ and $\mathcal{F}$ is a closed convex subset of $L_2(P_X^{(0)})$. Then, the transfer learning prediction function $\widehat{f}^{(0,tl)}$ in (11) under augmented responses as (9) has its excess risk satisfying that, with probability at least $1 - \delta$,

$$\begin{aligned}
R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \lesssim & \underbrace{\mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{N}}}_{\textbf{Standard generalization error}} + \underbrace{\frac{1}{N}\sum_{k=1}^{K} \|\widehat{w}_k - w_k\|_1}_{\textbf{Importance weight error}} \\
& + \underbrace{\frac{(N - n_0)^2}{N^2}\left[\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) + \frac{1}{M'}\right]}_{\textbf{Conditional transfer and Monte Carlo errors}}
\end{aligned}$$

where $\|\widehat{w}_k - w_k\|_1 = \sum_{i=1}^{n_k} \left|\widehat{w}_k(X_i^{(k)}) - w_k(X_i^{(k)})\right|$ and $\delta \in (0, 1)$.

If $\mathcal{F}$ is not convex, ℓ is a general loss function satisfying Assumption 3.1(a) only, and the augmented labels are independent calibrated draws $\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}$, then the last term becomes $\frac{N - n_0}{N}\sqrt{\widehat{\Delta}(\widehat{\boldsymbol{\beta}})}$.

Proof of Theorem 3.4. We first prove the excess risk bound under the squared loss $\ell(y, y') = (y-y')^2$ and $\mathcal{F}$ is a closed convex subset of $L_2(P_X^{(0)})$. Throughout the proof, we condition on $\left(\widehat{\boldsymbol{g}}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{w}}\right)$.

For (11), when $\widehat{Z}_i^{(k)} = \widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X_i^{(k)})$ in (9), we denote the empirical objective by

$$\widehat{R}_N^{\mu}(f; \widehat{\boldsymbol{w}}) := \frac{1}{N}\left\{\sum_{i=1}^{n_0}\left(Y_i^{(0)} - f(X_i^{(0)})\right)^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k}\widehat{w}_k(X_i^{(k)})\left(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X_i^{(k)}) - f(X_i^{(k)})\right)^2\right\},$$

and let $\widehat{R}_N^{\mu}(f; \boldsymbol{w})$ be the same objective with each $\widehat{w}_k$ replaced by the true density ratio w_k. We also define

$$R_{\widehat{\mu}}(f) = \mathbb{E}_{X\sim P_X^{(0)}, \boldsymbol{\eta}}\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - f(X)\right]^2, \quad R_A^{\mu}(f) = \frac{n_0}{N}\cdot R(f) + \left(\frac{N-n_0}{N}\right)R_{\widehat{\mu}}(f).$$

By the density-ratio identity under any function $h : \mathcal{X} \to \mathbb{R}$ as $\mathbb{E}_{P_X^{(k)}}[w_k(X)\cdot h(X)] = \mathbb{E}_{P_X^{(0)}}[h(X)]$, we know that

$$\mathbb{E}\left[\widehat{R}_N^{\mu}(f; \boldsymbol{w})\right] = R_A^{\mu}(f),$$

where the expectation is over $\left(X_i^{(0)}, Y_i^{(0)}\right)$ and $\left(X_i^{(k)}, Y_i^{(k)}\right), k = 1, ..., K$ as well as the independent noise vector $\boldsymbol{\eta} \sim P_\eta$.

Given the empirical minimizer $\widehat{f}^{(0,tl)}$, we have that

$$\begin{aligned} R_A^{\mu}(\widehat{f}^{(0,tl)}) - R_A^{\mu}(f^{(0)}) \le & \underbrace{\left[R_A^{\mu}(\widehat{f}^{(0,tl)}) - \widehat{R}_N^{\mu}(\widehat{f}^{(0,tl)}; \boldsymbol{w})\right] + \left[\widehat{R}_N^{\mu}(f^{(0)}; \boldsymbol{w}) - R_A^{\mu}(f^{(0)})\right]}_{\textbf{Term I}} \\ & + \underbrace{\left[\widehat{R}_N^{\mu}(\widehat{f}^{(0,tl)}; \boldsymbol{w}) - \widehat{R}_N^{\mu}(\widehat{f}^{(0,tl)}; \widehat{\boldsymbol{w}})\right] + \left[\widehat{R}_N^{\mu}(f^{(0)}; \widehat{\boldsymbol{w}}) - \widehat{R}_N^{\mu}(f^{(0)}; \boldsymbol{w})\right]}_{\textbf{Term II}}, \end{aligned} \tag{14}$$

where the inequality follows from the fact that $\widehat{R}_N^{\mu}(\widehat{f}^{(0,tl)}; \widehat{\boldsymbol{w}}) - \widehat{R}_N^{\mu}(f^{(0)}; \widehat{\boldsymbol{w}}) \le 0$.

• **Term I:** Similar to Proposition 3.3, we know that

$$\begin{aligned} \textbf{Term I} &\le 2\sup_{f\in\mathcal{F}}\left|\widehat{R}_N^{\mu}(f; \boldsymbol{w}) - R_A^{\mu}(f)\right| \\ &\le 4B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 2B_\ell B_w\sqrt{\frac{2\log(2/\delta)}{N}} \end{aligned}$$

with probability at least $1-\delta$.

• **Term II:** Uniformly over $f \in \mathcal{F}$, we know that

$$\begin{aligned} \left|\widehat{R}_N^{\mu}(f; \widehat{\boldsymbol{w}}) - \widehat{R}_N^{\mu}(f; \boldsymbol{w})\right| &= \left|\frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left[\widehat{w}_k(X_i^{(k)}) - w_k(X_i^{(k)})\right]\left(\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X_i^{(k)}) - f(X_i^{(k)})\right)^2\right| \\ &\le \frac{B_\ell}{N}\sum_{k=1}^{K}\sum_{i=1}^{n_k}\left|\widehat{w}_k(X_i^{(k)}) - w_k(X_i^{(k)})\right| \\ &= \frac{B_\ell}{N}\sum_{k=1}^{K}\left\|\widehat{w}_k - w_k\right\|_1, \end{aligned}$$

where ℓ satisfies Assumption 3.1(a) and $\left\|\widehat{w}_k - w_k\right\|_1 = \sum_{i=1}^{n_k}\left|\widehat{w}_k(X_i^{(k)}) - w_k(X_i^{(k)})\right|$ for $k = 1, ..., K$. Thus,

$$\textbf{Term II} \le \frac{2B_\ell}{N}\sum_{k=1}^{K}\left\|\widehat{w}_k - w_k\right\|_1.$$

Now, we connect $R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)})$ with $R(\widehat{f}^{(0,tl)}) - R(f^{(0)})$, which leads to the last "**conditional transfer and Monte Carlo errors**" term. Let $\mu^{(0)}(x) = \mathbb{E}(Y^{(0)} \mid X^{(0)} = x)$ and define

$$D_\mu^2 = \mathbb{E}_{X,\boldsymbol{\eta}}\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - \mu^{(0)}(X)\right]^2.$$

Under the squared loss $\ell(y, y') = (y - y')^2$,

$$R(f) = \mathbb{E}\left[\mathrm{Var}(Y^{(0)} \mid X^{(0)})\right] + \left\|\mu^{(0)} - f\right\|^2_{L_2(P_X^{(0)})}. \tag{15}$$

Thus, $f^{(0)} = \arg\min_{f\in\mathcal{F}} R(f)$ is the metric projection of $\mu^{(0)}$ onto the closed convex set $\mathcal{F}$. From (15), it follows that

$$\left\|f - f^{(0)}\right\|^2_{L_2(P_X^{(0)})} \le R(f) - R(f^{(0)}), \qquad \text{for every } f \in \mathcal{F}. \tag{16}$$

A direct expansion of the squared losses yields that

$$\begin{aligned}
&R_A^\mu(f) - R_A^\mu(f^{(0)}) \\
&= \frac{n_0}{N}\left[R(f) - R(f^{(0)})\right] + \left(\frac{N-n_0}{N}\right)\left\{\mathbb{E}_{X,\boldsymbol{\eta}}\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - f(X)\right]^2 - \mathbb{E}_{X,\boldsymbol{\eta}}\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - f^{(0)}(X)\right]^2\right\} \\
&= R(f) - R(f^{(0)}) - 2\left(\frac{N-n_0}{N}\right)\mathbb{E}_{X,\boldsymbol{\eta}}\left\{\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - \mu^{(0)}(X)\right]\left[f(X) - f^{(0)}(X)\right]\right\}.
\end{aligned}$$

Set $T = R(\widehat{f}^{(0,tl)}) - R(f^{(0)})$. Substituting $\widehat{f}^{(0,tl)}$ into the preceding display, we obtain that

$$\begin{aligned}
T &= R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \\
&= R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)}) + 2\left(\frac{N-n_0}{N}\right)\mathbb{E}_{X,\boldsymbol{\eta}}\left\{\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(X) - \mu^{(0)}(X)\right]\left[\widehat{f}^{(0,tl)}(X) - f^{(0)}(X)\right]\right\} \\
&\overset{\text{(i)}}{\le} R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)}) + 2\left(\frac{N-n_0}{N}\right) D_\mu \left\|\widehat{f}^{(0,tl)} - f^{(0)}\right\|_{L_2(P_X^{(0)})} \\
&\overset{\text{(ii)}}{\le} R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)}) + 2\left(\frac{N-n_0}{N}\right) D_\mu \sqrt{T} \\
&\overset{\text{(iii)}}{\le} R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)}) + \frac{T}{2} + 2\left(\frac{N-n_0}{N}\right)^2 D_\mu^2,
\end{aligned}$$

where (i) uses the Cauchy-Schwarz inequality, (ii) applies (16), and (iii) uses Young's inequality. Rearranging gives

$$\begin{aligned}
T &= R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \\
&\le 2\left[R_A^\mu(\widehat{f}^{(0,tl)}) - R_A^\mu(f^{(0)})\right] + 4\left(\frac{N-n_0}{N}\right)^2 D_\mu^2 \\
&\overset{\text{(iv)}}{\le} 8B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 4B_\ell B_w\sqrt{\frac{2\log(2/\delta)}{N}} + \frac{4B_\ell}{N}\sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 + 4\left(\frac{N-n_0}{N}\right)^2 D_\mu^2,
\end{aligned} \tag{17}$$

where (iv) follows from (14) and our results for **Term I** and **Term II**.

To control D_μ^2, we recall that $\widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x) = \mathbb{P}_{\boldsymbol{\eta}}\left[\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}(x,\boldsymbol{\eta}) \le t \middle| \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{g}}\right]$. Let $\bar{\mu}_{\widehat{\boldsymbol{\beta}}}(x) = \widehat{\beta}_0 + \sum_{k=1}^K \widehat{\beta}_k \cdot \mathbb{E}_\eta\left[\widehat{g}^{(k)}(x,\eta_k)\right]$. We know from (9) that

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\eta}}\left[\widehat{\mu}_{\widehat{\boldsymbol{\beta}}}(x) - \mu^{(0)}(x)\right]^2 &= \left[\bar{\mu}_{\widehat{\boldsymbol{\beta}}}(x) - \mu^{(0)}(x)\right]^2 + \frac{1}{M'}\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}(x,\boldsymbol{\eta})\middle|X = x\right] \\
&= \left\{\int_{\mathbb{R}}\left[F^{(0)}(t|x) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x)\right]dt\right\}^2 + \frac{1}{M'}\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}^T\widehat{\boldsymbol{V}}(x,\boldsymbol{\eta})\middle|X = x\right]
\end{aligned}$$

$$\overset{\text{(v)}}{\leq} 2B_g \int_{\mathbb{R}} \left[F^{(0)}(t|x) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x)\right]^2 dt + \frac{B_g^2}{M'},$$

where (v) follows from Assumption 3.2. Taking the expectation over $X \sim P^{(0)}$, we obtain that

$$D_\mu^2 \leq 2B_g \widehat{\Delta}(\widehat{\boldsymbol{\beta}}) + \frac{B_g^2}{M'}.$$

Substituting this inequality into (17) and noting that M' can be chosen to be sufficiently large, we derive that

$$\begin{aligned}
&R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) \\
&\lesssim 8B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 4B_\ell B_w \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{4B_\ell}{N} \sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 \\
&\quad + 4\left(\frac{N-n_0}{N}\right)^2 \left[2B_g \widehat{\Delta}(\widehat{\boldsymbol{\beta}}) + \frac{B_g^2}{M'}\right] \\
&\lesssim \mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{N}} + \frac{1}{N}\sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 + \left(\frac{N-n_0}{N}\right)^2 \left[\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) + \frac{1}{M'}\right].
\end{aligned}$$

This concludes the first part.

For the second part, suppose that $\mathcal{F}$ is not necessarily convex and that ℓ is a general loss function satisfying Assumption 3.1(a). When the augmented labels are sampled calibrated responses, we denote the empirical objective in (11) by

$$\widehat{R}_N^G(f; \widehat{\boldsymbol{w}}) := \frac{1}{N}\left\{\sum_{i=1}^{n_0} \ell\left(Y_i^{(0)}, f(X_i^{(0)})\right) + \sum_{k=1}^K \sum_{i=1}^{n_k} \widehat{w}_k(X_i^{(k)}) \cdot \ell\left(\widehat{\boldsymbol{\beta}}^T \widehat{\boldsymbol{V}}_i^{(k)}, f(X_i^{(k)})\right)\right\},$$

and let $\widehat{R}_N^G(f; \boldsymbol{w})$ be the same objective with the true density ratios. We also set

$$R_{\widehat{G}}(f) = \mathbb{E}_{X \sim P_X^{(0)}}\left\{\mathbb{E}_{Z \sim \widehat{G}_{\widehat{\boldsymbol{\beta}}}(\cdot|X)}\left[\ell(Z, f(X))\right]\right\}, \qquad R_A^G(f) = \frac{n_0}{N} \cdot R(f) + \left(\frac{N-n_0}{N}\right) R_{\widehat{G}}(f).$$

The density ratio identity again implies that $\mathbb{E}\left[\widehat{R}_N^G(f; \boldsymbol{w})\right] = R_A^G(f)$. Similar to (14) and the related **Term I** and **Term II**, we know that

$$R_A^G(\widehat{f}^{(0,tl)}) - R_A^G(f^{(0)}) \leq 4B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 2B_\ell B_w \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{2B_\ell}{N}\sum_{k=1}^K \|\widehat{w}_k - w_k\|_1. \tag{18}$$

Now, by the definition of R_A^G, we know that

$$\begin{aligned}
R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) &= R_A^G(\widehat{f}^{(0,tl)}) - R_A^G(f^{(0)}) + \left(\frac{N-n_0}{N}\right)\left[\left(R - R_{\widehat{G}}\right)(\widehat{f}^{(0,tl)}) - \left(R - R_{\widehat{G}}\right)(f^{(0)})\right] \\
&\leq 4B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 2B_\ell B_w \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{2B_\ell}{N}\sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 \\
&\quad + 2\left(\frac{N-n_0}{N}\right) \sup_{f \in \mathcal{F}} \left|R(f) - R_{\widehat{G}}(f)\right|.
\end{aligned} \tag{19}$$

Conditioning on $X = x$, we can couple $Y \sim F^{(0)}(\cdot|x)$ and $Z \sim \widehat{G}_{\widehat{\boldsymbol{\beta}}}(\cdot|x)$ by the common quantile coupling. By Assumption 3.1(a), we know that

$$\left|\mathbb{E}\left[\ell(Y, f(x))|X = x\right] - \mathbb{E}\left[\ell(Z, f(x))|X = x\right]\right| \leq B_\ell \int_{\mathbb{R}} \left|F^{(0)}(t|x) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x)\right| dt$$

$$\overset{\text{(vi)}}{\leq} B_\ell \sqrt{2B_g \int_{\mathbb{R}} \left[F^{(0)}(t|x) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x)\right]^2 dt},$$

where (vi) follows from Assumption 3.2 and the Cauchy-Schwarz inequality. Note that $\int_{\mathbb{R}} \left|F^{(0)}(t|x) - \widehat{G}_{\widehat{\boldsymbol{\beta}}}(t|x)\right| dt$ is the Wasserstein-1 distance between $F^{(0)}(\cdot|x)$ and $\widehat{G}_{\widehat{\boldsymbol{\beta}}}(\cdot|x)$.

Taking the expectation over $X \sim P_X^{(0)}$ and applying Jensen's inequality, we obtain that

$$\sup_{f \in \mathcal{F}} \left|R(f) - R_{\widehat{G}}(f)\right| \leq B_\ell \sqrt{2B_g \widehat{\Delta}(\widehat{\boldsymbol{\beta}})}.$$

Substituting this result into (19), we obtain that

$$\begin{aligned} R(\widehat{f}^{(0,tl)}) - R(f^{(0)}) &\leq 4B_\ell B_w \cdot \mathrm{Rad}_N(\mathcal{F}) + 2B_\ell B_w \sqrt{\frac{2\log(2/\delta)}{N}} + \frac{2B_\ell}{N} \sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 \\ &\quad + 2B_\ell \left(\frac{N - n_0}{N}\right) \sqrt{2B_g \widehat{\Delta}(\widehat{\boldsymbol{\beta}})} \\ &\lesssim \mathrm{Rad}_N(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{N}} + \frac{1}{N} \sum_{k=1}^K \|\widehat{w}_k - w_k\|_1 + \left(\frac{N - n_0}{N}\right) \sqrt{B_g \widehat{\Delta}(\widehat{\boldsymbol{\beta}})}. \end{aligned}$$

This completes the proof. □

E.3 Proof of Theorem 3.5

Theorem 3.5. Let $\delta \in (0,1)$ and $\left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty = \sup_{x \in \mathcal{X}} \mathbb{E}_\eta \left|\widehat{g}^{(k)}(x,\eta) - g^{(k)}(x,\eta)\right|$ for $k = 1, ..., K$. Suppose that Assumption 3.2 holds and that $\widehat{\boldsymbol{g}}$ is independent of the target observations used in (7). If $\widehat{\mathcal{C}}_{n_0,M}(\widehat{\boldsymbol{\beta}}) \leq \inf_{\boldsymbol{\beta} \in \Theta} \widehat{\mathcal{C}}_{n_0,M}(\boldsymbol{\beta}) + \epsilon_{\text{opt}}$ in (7), then with probability at least $1 - \delta$,

$$\begin{aligned} \widehat{\Delta}(\widehat{\boldsymbol{\beta}}) - \Delta(\boldsymbol{\beta}_*) &\lesssim \sum_{k=1}^K \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty + \sqrt{\frac{(\log n_0 + \log(1/\delta)) \left[\Delta(\boldsymbol{\beta}_*) + \sum_{k=1}^K \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty\right]}{n_0}} \\ &\quad + \frac{\log n_0 + \log(1/\delta)}{n_0} + \epsilon_{\text{opt}} \end{aligned}$$

up to some Monte Carlo approximation errors of order $O\left(\frac{1}{\sqrt{M}}\right)$.

Proof of Theorem 3.5. From (7), let

$$\widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta}) = \frac{1}{M} \sum_{j=1}^M \left|Y_i^{(0)} - \boldsymbol{\beta}^T \widehat{\boldsymbol{V}}_{ij}\right| - \frac{1}{2M(M-1)} \sum_{j,j'=1}^M \left|\boldsymbol{\beta}^T \widehat{\boldsymbol{V}}_{ij} - \boldsymbol{\beta}^T \widehat{\boldsymbol{V}}_{ij'}\right|.$$

Conditioning on $\widehat{\boldsymbol{g}}$, we know from (12) or Section E.1 and (5) that

$$\widehat{\Delta}(\boldsymbol{\beta}) = \mathbb{E}\left[\widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta}) \,\middle|\, \widehat{\boldsymbol{g}}\right] - \mathcal{C}_0$$

for any fixed $\boldsymbol{\beta} \in \Theta$. Then, we define

$$\begin{aligned} \widehat{D}_i(\boldsymbol{\beta}) &= \widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta}) - \mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X_i^{(0)}), Y_i^{(0)}\right) \\ &\quad + \mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X_i^{(0)}), Y_i^{(0)}\right) - \mathrm{CRPS}\left(F^{(0)}(\cdot|X_i^{(0)}), Y_i^{(0)}\right) \end{aligned}$$

Equivalently, $\widehat{D}_i(\boldsymbol{\beta}) = \widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta}) - \mathrm{CRPS}(F^{(0)}(\cdot|X_i^{(0)}), Y_i^{(0)})$. In addition, we have that

$$\begin{aligned} &\mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X), Y\right) - \mathrm{CRPS}\left(F^{(0)}(\cdot|X), Y\right) \\ &= \int_{\mathbb{R}} \left[\widehat{G}_{\boldsymbol{\beta}}(t|X) - F^{(0)}(t|X)\right] \left[\widehat{G}_{\boldsymbol{\beta}}(t|X) + F^{(0)}(t|X) - 2\mathbb{1}\{Y \leq t\}\right] dt. \end{aligned}$$

By the Cauchy-Schwarz inequality and Assumption 3.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left\{\left[\mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X), Y\right) - \mathrm{CRPS}\left(F^{(0)}(\cdot|X), Y\right)\right]^2 \Big| \widehat{\boldsymbol{g}}\right\} \\
&\overset{\text{(i)}}{\leq} 8B_g\, \mathbb{E}\left\{\int_{\mathbb{R}} \left[\widehat{G}_{\boldsymbol{\beta}}(t|X) - F^{(0)}(t|X)\right]^2 dt \,\Big|\, \widehat{\boldsymbol{g}}\right\} \\
&= 8B_g \widehat{\Delta}(\boldsymbol{\beta}),
\end{aligned}$$

where (i) uses the fact that Y is bounded by B_g, $\left|\widehat{G}_{\boldsymbol{\beta}}(t|X)\right| \leq 1$, and $\left|F^{(0)}(t|X)\right| \leq 1$. Moreover, by the Monte Carlo average and U-statistic structure,

$$\mathbb{E}\left[\widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta}) \,|\, \widehat{\boldsymbol{g}}, X_i^{(0)}, Y_i^{(0)}\right] = \mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X_i^{(0)}), Y_i^{(0)}\right),$$

and $\mathrm{Var}\left[\widehat{\mathcal{C}}_{i,M}(\boldsymbol{\beta})|\widehat{\boldsymbol{g}}, X_i^{(0)}, Y_i^{(0)}\right]$ is bounded by $\frac{C_1 B_g^2}{M}$ for a universal constant $C_1 > 0$. Therefore,

$$\mathrm{Var}\left[\widehat{D}_i(\boldsymbol{\beta}) \mid \widehat{\boldsymbol{g}}\right] \leq C_2 \left[\frac{B_g^2}{M} + B_g \widehat{\Delta}(\boldsymbol{\beta})\right]$$

for some universal constant $C_2 > 0$. Furthermore, under Assumption 3.2,

$$|\widehat{D}_i(\boldsymbol{\beta}) - \widehat{D}_i(\boldsymbol{\beta}')| \leq 2B_g \left\|\boldsymbol{\beta} - \boldsymbol{\beta}'\right\|_1$$

for any $\boldsymbol{\beta}, \boldsymbol{\beta}' \in \Theta$, and the same Lipschitz bound holds for $\widehat{\Delta}$. Since the L_1-ball Θ has an ϵ-net of cardinality at most $\left(1 + \frac{2B_\beta}{\epsilon}\right)^{K+1}$, Bernstein's inequality gives

$$\mathbb{P}\left(\left|\frac{1}{n_0}\sum_{i=1}^{n_0} \widehat{D}_i(\boldsymbol{\beta}) - \widehat{\Delta}(\boldsymbol{\beta})\right| \geq C_3 B_g \left\{\sqrt{\left[\frac{\widehat{\Delta}(\boldsymbol{\beta})}{B_g} + \frac{1}{M}\right]\frac{t}{n_0}} + \frac{t}{n_0}\right\}\right) \leq 2e^{-t}$$

for any $\boldsymbol{\beta} \in \Theta$. We now apply a union bound to a $\left(\frac{1}{n_0}\right)$-net and extend the result by Lipschitz continuity. Let $t_0(n_0, \delta) = (K+1)\log\left(1 + 2B_\beta n_0\right) + \log(2/\delta)$. Then, with probability at least $1 - \delta$, the following inequality holds simultaneously for every $\boldsymbol{\beta} \in \Theta$:

$$\begin{aligned}
\left|\frac{1}{n_0}\sum_{i=1}^{n_0} \widehat{D}_i(\boldsymbol{\beta}) - \widehat{\Delta}(\boldsymbol{\beta})\right| &\leq C_3 B_g \left\{\sqrt{\left[\frac{\widehat{\Delta}(\boldsymbol{\beta})}{B_g} + \frac{1}{M}\right]\frac{t_0(n_0,\delta)}{n_0}} + \frac{t_0(n_0,\delta)}{n_0}\right\} \\
&= C_3 B_g \left\{\sqrt{\left[\frac{\widehat{\Delta}(\boldsymbol{\beta})}{B_g} + \frac{1}{M}\right] a_{n_0}(\delta)} + a_{n_0}(\delta)\right\}
\end{aligned} \tag{20}$$

where $a_{n_0}(\delta) = \frac{t_0(n_0,\delta)}{n_0}$ and $C_3 > 0$ is an absolute constant.

Now, let $\bar{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}\in\Theta} \widehat{\Delta}(\boldsymbol{\beta})$, which exists by compactness of Θ. The empirical optimality of $\widehat{\boldsymbol{\beta}}$ and (20), together with Young's inequality to absorb the deviation term involving $\widehat{\Delta}(\widehat{\boldsymbol{\beta}})$, imply that

$$\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) \leq \inf_{\boldsymbol{\beta}\in\Theta} \widehat{\Delta}(\boldsymbol{\beta}) + CB_g \left\{\sqrt{\left[\frac{\inf_{\boldsymbol{\beta}\in\Theta}\widehat{\Delta}(\boldsymbol{\beta})}{B_g} + \frac{1}{M}\right] a_{n_0}(\delta)} + a_{n_0}(\delta)\right\} + C_4 \epsilon_{\mathrm{opt}} \tag{21}$$

with probability at least $1 - \delta$, where $C_4 > 0$ is an absolute constant.

Finally, we couple $\widehat{g}^{(k)}$ and $g^{(k)}$ with the same noise variables for $k = 1, ..., K$ and write $\widehat{Z}_{\boldsymbol{\beta}} - Z_{\boldsymbol{\beta}} = \sum_{k=1}^K \beta_k \left[\widehat{g}^{(k)}(X, \eta_k) - g^{(k)}(X, \eta_k)\right]$. By the triangle inequality,

$$\left|\mathbb{E}\left[\mathrm{CRPS}\left(\widehat{G}_{\boldsymbol{\beta}}(\cdot|X), Y\right) - \mathrm{CRPS}\left(G_{\boldsymbol{\beta}}(\cdot|X), Y\right)\right]\right|$$

$$
\begin{aligned}
&\leq \mathbb{E}\left|\mathbb{E}_{\boldsymbol{\eta}}|Y - Z_{\boldsymbol{\beta}}| - \mathbb{E}_{\boldsymbol{\eta}}|Y - \widehat{Z}_{\boldsymbol{\beta}}| - \frac{1}{2}\mathbb{E}_{\boldsymbol{\eta}}|Z_{\boldsymbol{\beta}} - Z'_{\boldsymbol{\beta}}| + \frac{1}{2}\mathbb{E}_{\boldsymbol{\eta}}|\widehat{Z}_{\boldsymbol{\beta}} - \widehat{Z}'_{\boldsymbol{\beta}}|\right| \\
&\leq 2B_\beta \sum_{k=1}^{K} \sup_{x\in\mathcal{X}} \mathbb{E}_\eta \left|\widehat{g}^{(k)}(x,\eta) - g^{(k)}(x,\eta)\right| \\
&= 2B_\beta \sum_{k=1}^{K} \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty
\end{aligned}
$$

uniformly over $\boldsymbol{\beta} \in \Theta$. It follows that

$$
\inf_{\boldsymbol{\beta}\in\Theta} \widehat{\Delta}(\boldsymbol{\beta}) \leq \underbrace{\inf_{\boldsymbol{\beta}\in\Theta} \Delta(\boldsymbol{\beta})}_{=\Delta(\boldsymbol{\beta}_*)} + 2B_\beta \sum_{k=1}^{K} \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty .
$$

Substituting this inequality into (21) and absorbing constants that depend only on B_g, B_β prove that

$$
\begin{aligned}
\widehat{\Delta}(\widehat{\boldsymbol{\beta}}) - \Delta(\boldsymbol{\beta}_*) \lesssim \sum_{k=1}^{K} \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty + \sqrt{\frac{(\log n_0 + \log(1/\delta)) \left[\Delta(\boldsymbol{\beta}_*) + \sum_{k=1}^{K} \left\|\widehat{g}^{(k)} - g^{(k)}\right\|_\infty\right]}{n_0}} \\
+ \frac{\log n_0 + \log(1/\delta)}{n_0} + \epsilon_{\text{opt}}
\end{aligned}
$$

up to some Monte Carlo approximation errors $O\left(\frac{1}{\sqrt{M}}\right)$. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the general distribution shift setting in (1), qualify the empirical claims by the size of the transfer bias, and point to the assumptions and theoretical guarantees of our proposed transfer learning framework in Section 3.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our work focuses on transfer regression learning with a univariate response variable. The proposed method relies on a shared linear calibration and may suffer negative transfer when its conditional family provides a poor approximation to the target distribution; see Remark 2.1. We examine this limitation empirically in Section C.3. Our theoretical results also require boundedness of the response variable and, for the downstream ERM bound, sample splitting or cross-fitting. We further discuss connections to related work in Section A, extensions to classification and multivariate responses in Section 5, and computational complexity in Section B.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions are stated in Section 3 before each result, and complete proofs are provided in Section E.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide code that reproduces our experiments in simulation studies and applications to a public dataset, along with implementation details in Section B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Code is provided in the supplement and on GitHub, and the public UCI dataset is openly available. The proprietary MAU data cannot be released.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are clearly delineated in Section B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The figures and tables report Monte Carlo standard errors across independent repetitions, and values below 0.001 are omitted from the MAU table.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section B.1 reports the compute type and cluster size, while the code specifies the memory allocated to each job.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: There are no ethical violations, to the authors' knowledge.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: An "Impact Statement" discussing potential benefits and risks is included before the references.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not provide any models or datasets in this paper. Our contributions are theoretical and methodological.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the UCI Apartment dataset and identifies it as licensed under CC BY 4.0. Meta Platforms, Inc. approved the publication of aggregate MAU evaluation results, but the underlying proprietary data are not redistributed.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released GitHub repository includes documentation for the code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Experiments of this nature are not included in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The experiments use publicly available tabular data and proprietary country-level aggregate statistics. They do not involve interaction with human subjects or collection of identifiable participant data. Meta Platforms, Inc. approved publication of the aggregate evaluation results.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No substantive part of this research involved the use of large language models.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.