
Wasserstein Transfer Learning

Kaicheng Zhang^{1*} Sinian Zhang^{2*} Doudou Zhou^{3†} Yidong Zhou^{4†}

¹School of Mathematical Sciences, Zhejiang University, China

²Division of Biostatistics and Health Data Science, University of Minnesota, USA

³Department of Statistics and Data Science, National University of Singapore, Singapore

⁴Department of Statistics, University of California, Davis, USA

3210102033@zju.edu.cn, zhan9381@umn.edu,

ddzhou@nus.edu.sg, ydzhou@ucdavis.edu

Abstract

Transfer learning is a powerful paradigm for leveraging knowledge from source domains to enhance learning in a target domain. However, traditional transfer learning approaches often focus on scalar or multivariate data within Euclidean spaces, limiting their applicability to complex data structures such as probability distributions. To address this limitation, we introduce a novel transfer learning framework for regression models whose outputs are probability distributions residing in the Wasserstein space. When the informative subset of transferable source domains is known, we propose an estimator with provable asymptotic convergence rates, quantifying the impact of domain similarity on transfer efficiency. For cases where the informative subset is unknown, we develop a data-driven transfer learning procedure designed to mitigate negative transfer. The proposed methods are supported by rigorous theoretical analysis and are validated through extensive simulations and real-world applications. The code is available at <https://github.com/h7nian/WaTL>.

1 Introduction

In recent years, transfer learning [32] has emerged as a powerful paradigm in machine learning, enabling models to leverage knowledge acquired from one domain and apply it to related tasks in another. This approach has proven especially valuable in scenarios where data collection and labeling can be costly, or where tasks exhibit inherent similarities in structure or representation. While early successes focused on conventional data types such as images [30], text [28], and tabular data [38], there is growing interest in extending these methods to more complex data structures. Such data often reside in non-Euclidean spaces and lack basic algebraic operations like addition, subtraction, or scalar multiplication, posing challenges for traditional learning algorithms. A key example is probability distributions [27], where for example the sum of two density functions does not yield a valid density.

Samples of univariate probability distributions are increasingly encountered across various research domains, such as mortality analysis [14], temperature studies [44], and physical activity monitoring [23], among others [27]. Recently, there has been a growing focus on directly modeling distributions as elements of the Wasserstein space, a geodesic metric space related to optimal transport [36, 24]. The absence of a linear structure in this space motivates the development of specialized transfer learning techniques that respect its intrinsic geometry.

To address this gap, we introduce *Wasserstein Transfer Learning* (WaTL), a novel transfer learning framework for regression models where outputs are univariate probability distributions. WaTL

*Equal contribution.

†Corresponding authors. Doudou Zhou and Yidong Zhou contributed equally.

effectively leverages knowledge from source domains to improve learning in a target domain by intrinsically incorporating the Wasserstein metric, which provides a natural way to measure discrepancies between probability distributions.

1.1 Contributions

The primary contributions of this work are summarized as follows:

Methodology. We propose a novel transfer learning framework for regression models with distributional outputs, addressing the challenges inherent in the Wasserstein space, which lacks a conventional linear structure. Our framework includes an efficient algorithm for cases where the informative subset of source domains is known, and a data-driven algorithm for scenarios where the subset is unknown. To the best of our knowledge, this is the first comprehensive transfer learning approach specifically designed for regression models with outputs residing in the Wasserstein space.

Theoretical analysis. We establish the asymptotic convergence rates for the WaTL algorithm in both the case where the informative set is known and the more challenging scenario where it must be estimated. In the latter case, we also prove that the informative set can be consistently identified. In both settings, we demonstrate that WaTL effectively improves model performance on the target data by leveraging information from the source domain. The proofs rely heavily on empirical process theory and a careful analysis of the covariate structure. Our key theoretical results extend beyond responses lying in the Wasserstein space, offering potential applications to other complex outputs.

Simulation studies and real-world applications. We evaluate WaTL through simulations and real data applications, demonstrating its effectiveness in improving target model performance by leveraging source domain information. The benefits become more pronounced with larger source sample sizes, underscoring its ability to harness transferable knowledge.

1.2 Related Work

Transfer learning. Transfer learning aims to improve performance in a target population by leveraging information from a related source population and has seen wide application across domains [e.g., 16, 15, 8, 41, 40]. Recent theoretical developments have focused on regression in Euclidean settings, including high-dimensional linear [21] and generalized linear models [31], nonparametric regression [6, 22], and function mean estimation from discretely sampled data [5]. In parallel, optimal transport has been used to measure distributional shifts for domain adaptation [10, 29]. However, to the best of our knowledge, no existing work has investigated transfer learning in regression models where outputs are probability distributions residing in the Wasserstein space. This represents a significant gap in the literature, highlighting the need for novel methodologies that address this challenging yet important setting.

Distributional data analysis. The increasing prevalence of data where distributions serve as fundamental units of observation has spurred the development of distributional data analysis [27]. Recent advancements in this field include geodesic principal component analysis in the Wasserstein space [1], autoregressive models for time series of distributions [39, 44], and distribution-on-distribution regression [14, 7]. Leveraging the Wasserstein metric, regression models with distributional outputs and Euclidean inputs can be viewed as a special case of Fréchet regression [26], which extends linear and local linear regression to outputs residing in general metric spaces. In practical scenarios where only finite samples from the unknown distributional output are available, empirical measures have been utilized as substitutes for the unobservable distributions in regression models [43].

2 Preliminaries

2.1 Notations

Let $L^2(0,1)$ be the space of square-integrable functions over the interval $(0,1)$, with the associated L^2 norm and metric denoted by $\|\cdot\|_2$ and d_{L^2} , respectively. To be specific, $\|g\|_2 = (\int_0^1 g^2(z)dz)^{1/2}$ and $d_{L^2}(g_1, g_2) = \|g_1 - g_2\|_2$. For a vector Z , $\|Z\|$ denotes the Euclidean norm. Given a matrix Σ , we define its spectrum as the set of its singular val-

ues. For a sub-Gaussian random vector X , we define the sub-Gaussian norm as $\|X\|_{\Psi_2} := \sup_{\|v\|=1} \inf \left\{ t > 0 : E(e^{\langle X, v \rangle^2 / t^2}) \leq 2 \right\}$.

We write $a_n \lesssim b_n$ if there exists a positive constant C such that $a_n \leq Cb_n$ when n is large enough and $a_n \asymp b_n$ if $a_n \lesssim b_n$ and $b_n \lesssim a_n$. The notation $a_n = O_p(b_n)$ implies that $P(|a_n/b_n| \leq C) \rightarrow 1$ for some constant $C > 0$, while $a_n = o_p(b_n)$ implies that $P(|a_n/b_n| > c) \rightarrow 0$ for any constant $c > 0$. Superscripts typically indicate different data sources, while subscripts distinguish individual samples from the same source.

2.2 Wasserstein Space

Let $\mathcal{W}$ denote the space of probability distributions on $\mathbb{R}$ with finite second moments, equipped with the 2-Wasserstein, or simply Wasserstein, metric. For two distributions $\mu_1, \mu_2 \in \mathcal{W}$, the Wasserstein metric is given by $d_{\mathcal{W}}^2(\mu_1, \mu_2) = \inf_{\pi \in \Pi(\mu_1, \mu_2)} \int_{\mathbb{R} \times \mathbb{R}} |s - t|^2 d\pi(s, t)$, where $\Pi(\mu_1, \mu_2)$ denotes the set of all joint distributions with marginals μ_1 and μ_2 [18]. For a probability measure $\mu \in \mathcal{W}$ with cumulative distribution function F_μ , we define the quantile function F_μ^{-1} as the left-continuous inverse of F_μ , such that $F_\mu^{-1}(u) = \inf\{t \in \mathbb{R} | F_\mu(t) \geq u\}$, for $u \in (0, 1)$. It has been established [36] that the Wasserstein metric can be expressed as the L^2 metric between quantile functions:

$$d_{\mathcal{W}}^2(\mu_1, \mu_2) = \int_0^1 \{F_{\mu_1}^{-1}(u) - F_{\mu_2}^{-1}(u)\}^2 du. \quad (1)$$

The space $\mathcal{W}$, endowed with the Wasserstein metric, forms a complete and separable metric space, commonly known as the Wasserstein space [36].

Assuming $E\{d_{\mathcal{W}}^2(\nu, \mu)\} < \infty$ for all $\mu \in \mathcal{W}$, the Fréchet mean [13] of a random distribution $\nu \in \mathcal{W}$ is given by $\nu_{\oplus} = \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu)\}$. Since the Wasserstein space $\mathcal{W}$ is a Hadamard space [20], the Fréchet mean is well-defined and unique. Moreover, from (1), it follows that the quantile function of the Fréchet mean, denoted as $F_{\nu_{\oplus}}^{-1}$, satisfies $F_{\nu_{\oplus}}^{-1}(u) = E\{F_{\nu}^{-1}(u)\}$, $u \in (0, 1)$.

2.3 Fréchet Regression

Consider a random pair (X, ν) with joint distribution $\mathcal{F}$ on the product space $\mathbb{R}^p \times \mathcal{W}$. Let X have mean $\theta = E(X)$ and covariance matrix $\Sigma = \text{Var}(X)$, where Σ is assumed to be positive definite. To establish a regression framework for predicting the distributional response ν from the covariate X , we employ the Fréchet regression model, which extends multiple linear regression and local linear regression to scenarios where responses reside in a metric space [26]. The Fréchet regression function is defined as the conditional Fréchet mean of ν given $X = x$,

$$m(x) = \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu) | X = x\}.$$

For a detailed exposition of Fréchet regression, we refer the reader to [26]. Given n independent realizations $\{(X_i, \nu_i)\}_{i=1}^n$, we define the empirical mean and covariance of X as $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ and $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^T$.

The global Fréchet regression extends classical multiple linear regression and estimates the conditional Fréchet mean as

$$m_G(x) = \arg \min_{\mu \in \mathcal{W}} E\{s_G(x) d_{\mathcal{W}}^2(\nu, \mu)\},$$

where the weight function is given by $s_G(x) = 1 + (X - \theta)^T \Sigma^{-1} (x - \theta)$. The empirical estimator is formulated as

$$\hat{m}_G(x) = \arg \min_{\mu \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n s_{iG}(x) d_{\mathcal{W}}^2(\nu_i, \mu),$$

where $s_{iG}(x) = 1 + (X_i - \bar{X})^T \hat{\Sigma}^{-1} (x - \bar{X})$.

Similarly, local Fréchet regression extends classical local linear regression to settings with metric space-valued outputs. In the case of a scalar predictor $X \in \mathbb{R}$, the local Fréchet regression function is

$$m_{L,h}(x) = \arg \min_{\mu \in \mathcal{W}} E\{s_L(x, h) d_{\mathcal{W}}^2(\nu, \mu)\},$$

where the weight function is $s_L(x, h) = K_h(X - x)\{u_2 - u_1(X - x)\}/\sigma_0^2$, with $u_j = E\{K_h(X - x)(X - x)^j\}$, $j = 0, 1, 2$, and $\sigma_0^2 = u_0u_2 - u_1^2$. Here, $K_h(\cdot) = h^{-1}K(\cdot/h)$ is a kernel function with bandwidth h . The empirical version is given by

$$\hat{m}_{L,h}(x) = \arg \min_{\mu \in \mathcal{W}} \frac{1}{n} \sum_{i=1}^n s_{iL}(x, h) d_{\mathcal{W}}^2(\nu_i, \mu),$$

where $s_{iL}(x, h) = K_h(X_i - x)\{\hat{u}_2 - \hat{u}_1(X_i - x)\}/\hat{\sigma}_0^2$, with $\hat{u}_j = n^{-1} \sum_{i=1}^n K_h(X_i - x)(X_i - x)^j$, $j = 0, 1, 2$, and $\hat{\sigma}_0^2 = \hat{u}_0\hat{u}_2 - \hat{u}_1^2$.

3 Methodology

3.1 Setup

We consider a transfer learning problem where target data $\{(X_i^{(0)}, \nu_i^{(0)})\}_{i=1}^{n_0}$ are sampled independently from the target population $(X^{(0)}, \nu^{(0)}) \sim \mathcal{F}_0$, and source data $\{(X_i^{(k)}, \nu_i^{(k)})\}_{i=1}^{n_k}$ are sampled independently from the source population $(X^{(k)}, \nu^{(k)}) \sim \mathcal{F}_k$, for $k = 1, \dots, K$. The goal is to estimate the target model using both the target data and source data from K related studies.

For $k = 0, \dots, K$, assume $X^{(k)}$ has mean θ_k and covariance Σ_k , with Σ_k positive definite. Define the empirical mean and covariance of $\{X_i^{(k)}\}_{i=1}^{n_k}$ as $\bar{X}_k = n_k^{-1} \sum_{i=1}^{n_k} X_i^{(k)}$ and $\hat{\Sigma}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} (X_i^{(k)} - \bar{X}_k)(X_i^{(k)} - \bar{X}_k)^\top$. For a fixed $x \in \mathbb{R}^p$, the weight function is $s_G^{(k)}(x) = 1 + (X^{(k)} - \theta_k)^\top \Sigma_k^{-1}(x - \theta_k)$, with the sample version $s_{iG}^{(k)}(x) = 1 + (X_i^{(k)} - \bar{X}_k)^\top \hat{\Sigma}_k^{-1}(x - \bar{X}_k)$. The target regression function for a given $x \in \mathbb{R}^p$ is then $m_G^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} E\{s_G^{(0)}(x) d_{\mathcal{W}}^2(\nu^{(0)}, \mu)\}$. In the following, we present details on transfer learning for global Fréchet regression, where the key difference in the local Fréchet regression setting is the use of a different weight function. The technical details for transfer learning in local Fréchet regression are therefore deferred to Appendix D.

The set of informative auxiliary samples (informative set) consists of sources sufficiently similar to the target data. Formally, the informative set is defined as $\mathcal{A}_\psi = \{1 \leq k \leq K : \|f^{(0)}(x) - f^{(k)}(x)\|_2 \leq \psi\}$ for some $\psi > 0$, where $f^{(k)}(x) = E\{s_G^{(k)}(x) F_{\nu^{(k)}}^{-1}\}$. For simplicity, let $n_{\mathcal{A}} = \sum_{k=1}^K n_k$.

3.2 Wasserstein Transfer Learning

We propose the Wasserstein Transfer Learning (WaTL) algorithm, which combines information from source datasets under the assumption that all source data are *informative enough*. This assumption implies that the discrepancies between the source and target are small enough to enhance estimation compared to using only the target. When this condition is met for all source datasets, the informative set is given by $\mathcal{A}_\psi = \{1, \dots, K\}$. The detailed steps of WaTL are presented in Algorithm 1.

Algorithm 1 Wasserstein Transfer Learning (WaTL)

Input: Target and source data $\{(x_i^{(0)}, \nu_i^{(0)})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \nu_i^{(k)})\}_{i=1}^{n_k})$, regularization parameter λ , and query point $x \in \mathbb{R}^p$.

Output: Target estimator $\hat{m}_G^{(0)}(x)$.

- 1: Weighted auxiliary estimator: $\hat{f}(x) = \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K n_k \hat{f}^{(k)}(x)$, where $\hat{f}^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) F_{\nu_i^{(k)}}^{-1}$.
 - 2: Bias correction using target data: $\hat{f}_0(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) \|F_{\nu_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}(x)\|_2$.
 - 3: Projection to Wasserstein space: $\hat{m}_G^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_0(x)\|_2$.
-

In Step 1, the initial estimate $\hat{f}$ aggregates information from both the target and source, weighted by their respective sample sizes. While this step incorporates valuable auxiliary information, the resulting

Algorithm 2 Adaptive Wasserstein Transfer Learning (AWaTL)

Input: Target and source data $\{(x_i^{(0)}, \nu_i^{(0)})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \nu_i^{(k)})\}_{i=1}^{n_k})$, regularization parameter λ , prespecified number of informative sources L , and query point $x \in \mathbb{R}^p$.

Output: Target estimator $\hat{m}_G^{(0)}(x)$.

- 1: Compute discrepancy scores. For each source dataset $k = 1, \dots, K$, compute the empirical discrepancy: $\hat{\psi}_k = \|\hat{f}^{(0)}(x) - \hat{f}^{(k)}(x)\|_2$, where $\hat{f}^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) F_{\nu_i^{(k)}}^{-1}$. Construct the adaptive informative set by selecting the L smallest discrepancy scores $\hat{\mathcal{A}} = \{k : 1 \leq k \leq K \text{ and } \hat{\psi}_k \text{ is among the } L \text{ smallest values}\}$.
 - 2: Weighted auxiliary estimator: $\hat{f}(x) = \frac{1}{\sum_{k \in \hat{\mathcal{A}} \cup \{0\}} n_k} \sum_{k \in \hat{\mathcal{A}} \cup \{0\}} n_k \hat{f}^{(k)}(x)$.
 - 3: Bias correction using target data: $\hat{f}_0(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) \|F_{\nu_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}(x)\|_2$.
 - 4: Projection to Wasserstein space: $\hat{m}_G^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_0(x)\|_2$.
-

estimate may be biased due to distributional differences between the target and source populations. In Step 2, the bias in $\hat{f}$ is corrected by focusing on the target data. The regularization term $\lambda \|g - \hat{f}(x)\|_2$ ensures a balance between target-specific precision and auxiliary-informed robustness. Theoretical guidelines for selecting λ are provided in Theorem 2. The final step projects the corrected estimate $\hat{f}_0$ onto the Wasserstein space, ensuring the output $\hat{m}_G^{(0)}(x)$ respects the intrinsic geometry of $\mathcal{W}$. This projection exists and is unique because $\mathcal{W}$ is a closed and convex subset of $L^2(0, 1)$.

3.3 Adaptive Selection of Informative Sources

In many practical scenarios, the assumption that all source datasets belong to the informative set $\mathcal{A}_\psi$ may not hold. To address this, we extend WaTL with an adaptive selection procedure to identify the informative set. The discrepancy for each source dataset k is defined as $\psi_k = \|f^{(0)}(x) - f^{(k)}(x)\|_2$, which measures the distance between the target distribution and the auxiliary distribution. Since $f^{(0)}(x)$ and $f^{(k)}(x)$ are unknown, we compute an empirical estimate $\hat{\psi}_k$ for ψ_k , which is used to adaptively estimate the informative set $\mathcal{A}_\psi$. To implement this approach, an additional input parameter L , which specifies the approximate number of informative sources, is required. In practice, L can be treated as a tuning parameter and selected through cross-validation or other model selection techniques. The full procedure is formalized in Algorithm 2.

The proposed algorithm adaptively identifies the informative set $\hat{\mathcal{A}}$ in Step 1 by evaluating the empirical discrepancy scores $\hat{\psi}_k$. The selected set is then used to compute the weighted auxiliary estimator in Step 2, ensuring that only the most relevant source datasets contribute to the final target estimator. Steps 3 and 4 follow the same bias correction and projection procedures as described in Section 3.2. This adaptive approach enhances the robustness of WaTL by excluding irrelevant or highly dissimilar source datasets.

4 Theory

In this section, we establish the theoretical guarantees of the proposed WaTL and AWaTL algorithms using techniques from empirical process theory [34]. For WaTL, we present the following lemma, which characterizes the convergence rate for each term contributing to the weighted auxiliary estimator $\hat{f}(x)$ computed in Step 1.

Condition 1. For $k = 0, \dots, K$, the covariate $X^{(k)}$ is sub-Gaussian with $\|X^{(k)}\|_{\Psi_2} \in [\sigma_1, \sigma_2]$, the mean vector satisfies $\|\theta_k\| \leq R_1$, and the spectrum of the covariance matrix Σ_k lies within the interval $[R_2, R_3]$. Moreover, $\nu^{(k)}$ is supported on a bounded interval.

Lemma 1. Let $\hat{f}^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) F_{\nu_i^{(k)}}^{-1}$ and its population counterpart be defined as $f^{(k)}(x) = E\{s_G^{(k)}(x) F_{\nu^{(k)}}^{-1}\}$ for $k = 0, \dots, K$. Then under Condition 1, $\|\hat{f}^{(k)}(x) - f^{(k)}(x)\|_2 = O_p(n_k^{-1/2})$.

To derive the convergence rate of $\hat{f}(x)$, we rely on the following condition.

Condition 2. There exist positive constants C_1, C_2, C_3, C_4 such that

$$\sum_{k=0}^K 2e^{-C_1 \frac{4}{R_3^2} n_k} + C_2 e^{-C_3 \left(\frac{2}{R_3} - \sqrt{\frac{C_4}{n_k}} \right) n_k} = o(1),$$

where R_3 is as in Condition 1. In addition,

$$\frac{\sqrt{n_0 + n_A}}{\min_{1 \leq k \leq K} n_k} = o(1), \quad \frac{\sqrt{n_0 + n_A}}{n_0} = o(1).$$

Remark 1. These conditions are typically satisfied in practice as they are not overly restrictive. Condition 1 requires that covariates and covariance matrices are bounded in a specific way, which is standard in the transfer learning literature and generally holds in real-world scenarios [5]. In particular, this assumption is common when dealing with high-dimensional data where regularization is necessary [21]. Condition 2 assumes that the number of samples is significantly larger than the number of sources K , which is reasonable since K is usually fixed in practical settings, and we often have sufficient source data compared to target data. In practice, Condition 2 may be slightly violated if there exists a source k with a relatively small n_k , such that $\sqrt{n_0 + n_A}/n_k$ is not $o(1)$. In such cases, the k th source can simply be excluded from Step 1 of the WaTL algorithm.

Theorem 1. Suppose Conditions 1 and 2 hold. Then, for the WaTL algorithm, it holds for a fixed $x \in \mathbb{R}^p$ that

$$\|\hat{f}(x) - f(x)\|_2 = O_p\left(\frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_A} + (n_0 + n_A)^{-1/2}\right),$$

where $f(x) = (n_0 + n_A)^{-1} \sum_{k=0}^K n_k f^{(k)}(x)$.

The proof of Theorem 1 involves a detailed analysis of the sample covariance matrix and leverages M-estimation techniques within the framework of empirical process theory. The result extends beyond responses in the Wasserstein space, applying to other metric spaces that meet mild regularity conditions. Consequently, Theorem 1 provides a versatile framework that can be applied to transfer learning in regression models with responses such as networks [42], symmetric positive-definite matrices [25], or trees [2]. The following theorem establishes the convergence rate for the estimated regression function $\hat{m}_G^{(0)}(x)$ in the WaTL algorithm.

Theorem 2. Assume Conditions 1 and 2 hold and the regularization parameter satisfies $\lambda \asymp n_0^{-1/2+\epsilon}$ for some $\epsilon > 0$. Then, for the WaTL algorithm and a fixed $x \in \mathbb{R}^p$, it holds that

$$d_W^2(\hat{m}_G^{(0)}(x), m_G^{(0)}(x)) = O_p\left(n_0^{-1/2+\epsilon} \left(\psi + \frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_A} + (n_0 + n_A)^{-1/2}\right)\right),$$

where $\psi = \max_{1 \leq k \leq K} \|f^{(0)}(x) - f^{(k)}(x)\|_2$ quantifies the maximum discrepancy between the target and source.

Theorem 2 can be compared to the convergence rate of global Fréchet regression [26] applied solely to the target data, for which the rate is $d_W(\hat{m}_G^{(0)}(x), m_G^{(0)}(x)) = O_p(n_0^{-1/2})$. The WaTL algorithm achieves a faster convergence rate when there are sufficient source data and the auxiliary sources are informative enough, satisfying $\psi \lesssim n_0^{-1/2-\epsilon}$. This result highlights that knowledge transfer from auxiliary samples can significantly enhance the learning performance of the target model, provided the auxiliary sources are closely aligned with the target data.

For the AWaTL algorithm, we require the following condition.

Condition 3. The regularization parameter satisfies $\lambda \asymp n_0^{-1/2+\epsilon}$ for some $\epsilon > 0$ and for some $\epsilon' > \epsilon$, there exists a non-empty subset $\mathcal{A} \subset \{1, \dots, K\}$ such that

$$\frac{n_*^{-1/2} + n_0^{-1/2}}{\min_{k \in \mathcal{A}^c} \psi_k} = o(1), \quad \frac{\max_{k \in \mathcal{A}} \psi_k}{n_*^{-1/2} + n_0^{-1/2-\epsilon'}} = O(1),$$

where $\mathcal{A}^C = \{1, \dots, K\} \setminus \mathcal{A}$, $n_* = \min_{1 \leq k \leq K} n_k$ and $\psi_k = \|f^{(0)}(x) - f^{(k)}(x)\|_2$.

Remark 2. We allow $\mathcal{A}$ to be the whole set $\{1, \dots, K\}$, in which case the condition becomes

$$\frac{\max_{1 \leq k \leq K} \psi_k}{n_*^{-1/2} + n_0^{-1/2-\epsilon'}} = O(1).$$

Besides, if $\mathcal{A}$ exists, then it is unique since

$$\{1 \leq k \leq K : \frac{n_*^{-1/2} + n_0^{-1/2}}{\psi_k} = o(1)\} \cap \{1 \leq k \leq K : \frac{\psi_k}{n_*^{-1/2} + n_0^{-1/2-\epsilon'}} = O(1)\} = \emptyset.$$

Condition 3 ensures that the source datasets can be effectively partitioned into two groups: informative ones and those sufficiently different from the target data. This separation guarantees that the informative set can be accurately identified. Under this condition, by setting the number of informative sources to $L = |\mathcal{A}|$, we establish the following rate of convergence for the AWaTL algorithm. In practice, L can be decided by cross-validation.

Theorem 3. *Under Condition 3 and the conditions of Theorem 2, for the AWaTL algorithm with a fixed number of sources K we have $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$ and*

$$d_{\mathcal{W}}^2(\hat{m}_G^{(0)}(x), m_G^{(0)}(x)) = O_p\left(n_0^{-1/2+\epsilon} \left(\max_{k \in \mathcal{A}} \psi_k + \frac{\sum_{k \in \mathcal{A} \cup \{0\}} \sqrt{n_k}}{\sum_{k \in \mathcal{A} \cup \{0\}} n_k} + \left(\sum_{k \in \mathcal{A} \cup \{0\}} n_k \right)^{-1/2} \right)\right).$$

The convergence rate of the AWaTL algorithm simplifies to $n_0^{-1-\epsilon'+\epsilon}$ when the informative source data is sufficiently large. This rate surpasses that of global Fréchet regression [26] applied solely to the target data, offering a theoretical guarantee that AWaTL effectively mitigates negative transfer by selectively integrating relevant auxiliary information.

While the above analysis assumes that each probability distribution is fully observed, an assumption commonly adopted in the distributional data literature [26, 7], real-world applications often provide only independent samples drawn from underlying distributions. In such cases, this limitation can be overcome by replacing unobservable distributions with their empirical counterparts, constructed from sample observations [43]. Additional details and theoretical justification for this extension are provided in Appendix E.

5 Numerical Experiments

In this section, we evaluate the performance of the proposed WaTL algorithm, alongside two baseline approaches: the global Fréchet regression using only target data (Only Target) and using only source data (Only Source). Consider $K = 5$ source sites. The data are generated as follows. For the target population, we sample $X^{(0)} \sim U(0, 1)$ and generate the response distribution, represented by its quantile function, as

$$F_{\nu^{(0)}}^{-1}(u) = w^{(0)}(1-u)u + (1-X^{(0)})u + X^{(0)}F_{Z^{(0)}}^{-1}(u), \quad u \in (0, 1),$$

where $Z^{(0)} \sim N(0.5, 1)|_{(0,1)}$ and $w^{(0)} \sim N(0, 1)|_{(-0.5, 0.5)}$. Here, $N(\mu, \sigma^2)|_{(a,b)}$ denotes a normal distribution with mean μ and variance σ^2 , truncated to the interval (a, b) . For source populations, we define $\psi_k = 0.1k$ for $k = 1, \dots, K$, and generate $X^{(k)} \sim U(0, 1)$. The corresponding response distribution is generated as

$$F_{\nu^{(k)}}^{-1}(u) = w^{(k)}(1-u)u + (1-X^{(k)})u + X^{(k)}F_{Z^{(k)}}^{-1}(u), \quad u \in (0, 1),$$

where $Z^{(k)} \sim N(0.5, 1-\psi_k)|_{(0,1)}$ and $w^{(k)} \sim N(0, 1)|_{(-0.5, 0.5)}$. Consequently, for each predictor x , the true regression function is $m_G^{(k)}(x) = (1-x)u + xF_{Z^{(k)}}^{-1}(u)$, for $k = 0, 1, \dots, K$.

We vary the target sample size n_0 from 200 to 800, while the source sample size is set as $n_k = k\tau$, where $\tau \in \{100, 200\}$ and $k = 1, \dots, K$. The regularization parameter λ in Algorithm 1 is selected via five-fold cross-validation, ranging from 0 to 3 in increments of 0.1. To evaluate performance, we sample 100 predictors uniformly from the target distribution. Using Algorithm 1, we compute

$\hat{m}_G^{(0)}(x)$ and compare it with the corresponding estimates obtained from global Fréchet regression using only target or source data. Performance is assessed using the root mean squared prediction risk $\text{RMSPR} = \sqrt{\frac{1}{100} \sum_{i=1}^{100} d_W^2(\hat{m}_G^{(0)}(x_i), m_G^{(0)}(x_i))}$, where x_i denotes the sampled predictor, $\hat{m}_G^{(0)}(x_i)$ is the estimated function, and $m_G^{(0)}(x_i)$ represents the ground truth. To ensure robustness, we repeat the simulation 50 times and report the average RMSPR, as shown in Figure 1(a).

As shown in Figure 1(a), WaTL consistently outperforms global Fréchet regression trained solely on target or source data. When the target sample size n_0 is small, the Only Target method exhibits a high RMSPR due to the instability of models trained on limited data. In contrast, WaTL significantly reduces RMSPR by effectively incorporating auxiliary information from the source domain. As n_0 increases, the performance of Only Target improves and gradually approaches that of WaTL, which is expected as larger sample sizes lead to more stable and accurate estimators. Nevertheless, WaTL maintains a consistent advantage across all n_0 , suggesting that it leverages complementary information from the source. The performance of the Only Source estimator remains nearly unchanged across different n_0 values, as it does not benefit from additional target data.

Comparing the two panels of Figure 1(a), we also observe that WaTL improves as the source sample size increases, confirming its ability to effectively integrate information from source domains. This demonstrates the benefit of multi-source transfer learning, where WaTL balances knowledge from both target and source domains to achieve improved prediction.

To better understand when negative transfer may occur, we conduct an ablation study with $K = 1$ source and vary the similarity parameter ψ_1 from 0.01 to 1 in increments of 0.01, with $n_0 = 100$ and $n_1 = 200$. Our method outperforms the Only Target approach when $\psi_1 < 0.9$, while for $\psi_1 \geq 0.9$, the Only Target method becomes preferable. This confirms that negative transfer arises when the source is too dissimilar to the target.

We further evaluate the effectiveness of AWaTL in selecting informative sources. In this experiment, we set $L = 2$ and $n_k = 100$ for $k = 0, \dots, 5$. The similarity parameters are specified as $\psi_k = 0.1$ for $k = 1, 2$ (informative sources), and $\psi_k = \psi$, increasing from 0.2 to 1 in increments of 0.1, for $k = 3, 4, 5$ (uninformative sources). Each configuration is repeated 100 times, and the corresponding selection rates are reported in Figure 1(b). The results show that AWaTL successfully identifies informative sources, with selection rates for sources 1 and 2 rapidly increasing and reaching perfect accuracy once $\psi > 0.6$. These findings demonstrate the robustness of AWaTL in distinguishing useful sources under varying similarity levels.

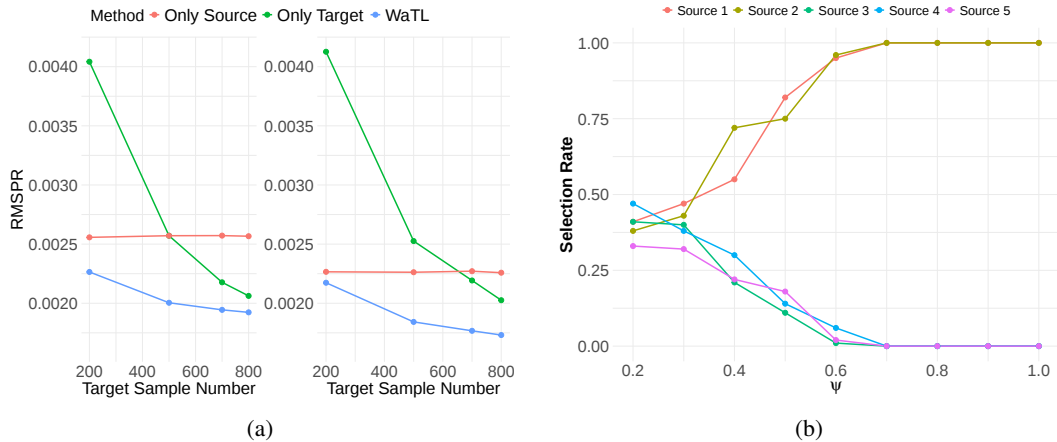


Figure 1: (a) Root mean squared prediction risk (RMSPR) of WaTL, only Source, and Only Target methods under varying target sample sizes, with source sample sizes $\tau = 100$ (left) and $\tau = 200$ (right); (b) Selection rate of each source site as ψ increases.

6 Real-world Applications

We evaluate the WaTL algorithm using data from the National Health and Nutrition Examination Survey (NHANES) 2005–2006³, focusing on modeling the distribution of physical activity intensity. NHANES is a large-scale health survey in the United States that combines interviews with physical examinations to assess the health and nutrition of both adults and children. The dataset includes extensive demographic, socioeconomic, dietary, and medical assessments, providing a comprehensive resource for health-related research.

During the 2005–2006 NHANES cycle, participants aged 6 and older wore an ActiGraph 7164 accelerometer on their right hip for seven days, recording physical activity intensity in 1-minute epochs. Participants were instructed to remove the device during water-based activities and sleep. The device measured counts per minute (CPM), ranging from 0 to 32767, capturing variations in activity levels throughout the monitoring period.

Since female and male participants exhibit distinct physical activity patterns [12], we analyze them separately. Physical activity intensity is influenced by multiple factors, and we consider body mass index (BMI) and age as key predictors [19, 9]. To accommodate potential nonlinear relationships, we implement local Fréchet regression within the WaTL algorithm, treating the distribution of physical activity intensity as the response and BMI and age as predictors.

Following the data preprocessing steps in [23], we remove unreliable observations per NHANES protocols. For each participant, we exclude activity counts above 1000 CPM or equal to zero (as zeros may correspond to various low-activity states such as sleep or swimming). Participants with fewer than 100 valid observations or missing BMI, age, or gender information are also excluded. The remaining activity counts over seven days are concatenated to form the distribution of each participant’s activity intensity.

To evaluate the WaTL, we set White, Mexican Americans, and other Hispanic individuals as sources and Black as the target. For females, the source data include 1308 White people, 884 Mexican Americans, and 108 Other Hispanic individuals. For males, the source data include 1232 White participants, 805 Mexican Americans, and 92 Other Hispanic individuals. We set 200 Black participants as the target data for both genders. During evaluating, we perform five-fold cross-validation, using four folds for training and one for testing, cycling through each fold. For comparison, we also apply local Fréchet regression using only the target data. The results, summarized in Figure 2(a), show that WaTL improves performance over local Fréchet regression for both females and males, demonstrating its ability to leverage information from other demographic groups to enhance modeling for Black participants.

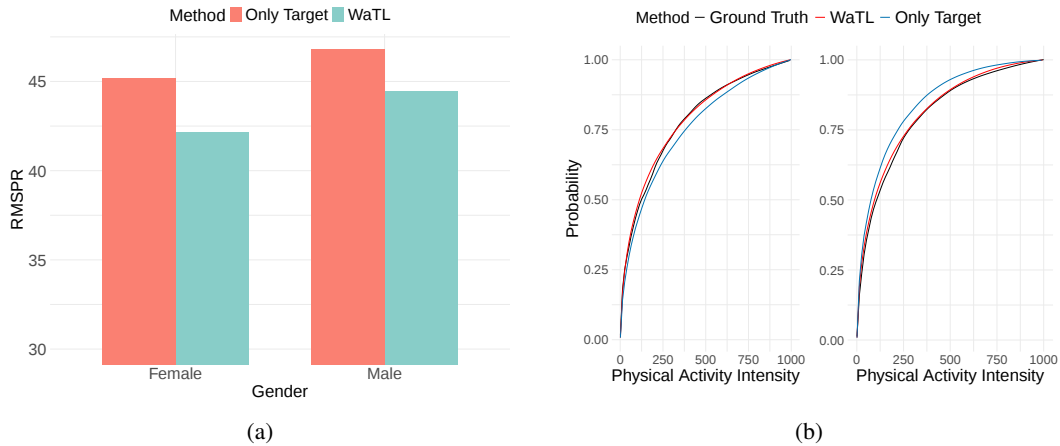


Figure 2: (a) Root mean squared prediction risk (RMSPR) of WaTL and Only Target methods for females and males, evaluated using five-fold cross-validation; (b) Cumulative distribution function of physical activity levels for one selected female (left) and one selected male (right), along with estimates from WaTL and Only Target methods.

³<https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2005>

To further illustrate the effectiveness of WaTL, we visualize the cumulative distribution function of physical activity levels for one female participant and one male participant in Figure 2(b), along with estimates from WaTL and local Fréchet regression, using only the target data. The results indicate that WaTL provides a better fit to the true distribution, outperforming the estimate obtained using only the target data. We further evaluate the robustness of WaTL on a human mortality dataset in Appendix A, where WaTL continues to demonstrate strong performance.

7 Conclusion

We introduce Wasserstein transfer learning and its adaptive variant for scenarios where the informative set is unknown, addressing the challenges posed by the lack of linear operations in the Wasserstein space. By leveraging the Wasserstein metric, the proposed algorithm accounts for the non-Euclidean structure of distributional outputs, ensuring compatibility with the intrinsic geometry of the Wasserstein space. Supported by rigorous theoretical guarantees, the framework demonstrates improved estimation performance compared to methods that rely solely on target data.

This paper focuses on univariate distributions, which arise frequently in real-world applications such as the NHANES and human mortality studies discussed in Section 6. The proposed framework, however, is not limited to the univariate case and can be extended to multivariate distributions by incorporating metrics such as the Sinkhorn [11] or sliced Wasserstein distance [4]. Extending the methodology to higher dimensions is an exciting future direction that entails addressing both computational and theoretical challenges specific to high-dimensional Wasserstein spaces.

Acknowledgments and Disclosure of Funding

We thank the Area Chair and the reviewers for their constructive feedback. Doudou Zhou was supported by the NUS Start-Up Grant (A-0009985-00-00) and the MOE AcRF Tier 1 Grant (A-8003569-00-00).

References

- [1] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré B: Probability and Statistics*, 53:1–26, 2017.
- [2] Louis J. Billera, Susan P. Holmes, and Karen Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.
- [3] Sergey Bobkov and Michel Ledoux. *One-dimensional Empirical Measures, Order Statistics, and Kantorovich Transport Distances*, volume 261. American Mathematical Society, 2019.
- [4] 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.
- [5] T. Tony Cai, Dongwoo Kim, and Hongming Pu. Transfer learning for functional mean estimation: Phase transition and adaptive algorithms. *Annals of Statistics*, 52(2):654 – 678, 2024.
- [6] T Tony Cai and Hongming Pu. Transfer learning for nonparametric regression: Non-asymptotic minimax analysis and adaptive procedure. *arXiv preprint arXiv:2401.12272*, 2024.
- [7] Yaqing Chen, Zhenhua Lin, and Hans-Georg Müller. Wasserstein regression. *Journal of the American Statistical Association*, 118(542):869–882, 2023.
- [8] Keunwoo Choi, George Fazekas, Mark Sandler, and Kyunghyun Cho. Transfer learning for music classification and regression tasks. In *The 18th International Society of Music Information Retrieval (ISMIR) Conference 2017, Suzhou, China*. International Society of Music Information Retrieval, 2017.

- [9] Laura Cleven, Jeremy A Syrjanen, Yonas E Geda, Luke R Christenson, Ronald C Petersen, Maria Vassilaki, Alexander Woll, and Janina Krell-Roesch. Association between physical activity and longitudinal change in body mass index in middle-aged and older adults. *BMC Public Health*, 23(1):202, 2023.
- [10] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Advances in Neural Information Processing Systems*, volume 30, pages 3730–3739, 2017.
- [11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [12] Alessandra Feraco, Stefania Gorini, Elisabetta Camajani, Tiziana Filardi, Serkan Karav, Edda Cava, Rocky Strollo, Elvira Padua, Massimiliano Caprio, Andrea Armani, and Mauro Lombardo. Gender differences in dietary patterns and physical activity: An insight with principal component analysis (PCA). *Journal of Translational Medicine*, 22(1):1112, 2024.
- [13] Maurice Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l’Institut Henri Poincaré*, 10(4):215–310, 1948.
- [14] Laya Ghodrati and Victor M Panaretos. Distribution-on-distribution regression via optimal transport maps. *Biometrika*, 109(4):957–974, 2022.
- [15] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2066–2073, 2012.
- [16] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308, 2013.
- [17] Su I Iao, Yidong Zhou, and Hans-Georg Müller. Deep Fréchet regression. *Journal of the American Statistical Association*, 2025. in press.
- [18] Leonid V Kantorovich. On the translocation of masses. *Dokl. Akad. Nauk SSSR (translated version in Journal of Mathematical Sciences, 133, 1381-1382, 2006)*, 37:227–229, 1942.
- [19] Thomas Klein and Simone Becker. Age and exercise: a theoretical and empirical analysis of the effect of age and generation on physical activity. *Journal of Public Health*, 20:11–21, 2012.
- [20] Benoit Kloeckner. A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 9(2):297–323, 2010.
- [21] 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.
- [22] 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.
- [23] Zhenhua Lin, Dehan Kong, and Linbo Wang. Causal inference on distribution functions. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(2):378–398, 2023.
- [24] Victor M Panaretos and Yoav Zemel. *An Invitation to Statistics in Wasserstein Space*. Springer New York, 2020.
- [25] Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.
- [26] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics*, 47(2):691–719, 2019.

- [27] Alexander Petersen, Chao Zhang, and Piotr Kokoszka. Modeling probability density functions as data objects. *Econometrics and Statistics*, 21:159–178, 2022.
- [28] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [29] Ievgen Redko, Amaury Habrard, and Marc Sebban. Theoretical analysis of domain adaptation with optimal transport. In Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski, Celine Vens, and Sašo Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 737–753, Cham, 2017. Springer International Publishing.
- [30] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.
- [31] Ye Tian and Yang Feng. Transfer learning under high-dimensional generalized linear models. *Journal of the American Statistical Association*, 118(544):2684–2697, 2023.
- [32] 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, 2010.
- [33] Aad W. van der Vaart. *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press, 2000.
- [34] Aad W. van der Vaart and John Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer New York, 2nd edition, 2023.
- [35] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.
- [36] Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.
- [37] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.
- [38] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(9):1–40, 2016.
- [39] Chao Zhang, Piotr Kokoszka, and Alexander Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 43(2):30–52, 2022.
- [40] 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.
- [41] 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*, pages 1–26, 2024.
- [42] Yidong Zhou and Hans-Georg Müller. Network regression with graph Laplacians. *Journal of Machine Learning Research*, 23(320):1–41, 2022.
- [43] Yidong Zhou and Hans-Georg Müller. Wasserstein regression with empirical measures and density estimation for sparse data. *Biometrics*, 80(4):ujae127, 2024.
- [44] Changbo Zhu and Hans-Georg Müller. Autoregressive optimal transport models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):1012–1033, 2023.

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 accurately state the paper's contributions, including the proposal of the WaTL framework, theoretical analysis of convergence rates, and validation through simulations and real-world data. These are detailed in subsequent sections (e.g., Section 1.1, Section 3, Section 4).

Guidelines:

- The answer NA 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 NA 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: The paper discusses a limitation concerning the focus on univariate distributions and suggests how this could be addressed in practical settings. This is found in Section 7.

Guidelines:

- The answer NA 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: Theoretical results (Lemmas and Theorems) are presented in Section 4, with assumptions clearly stated (Conditions 1–3). The paper mentions that detailed proofs are provided in Appendix C.

Guidelines:

- The answer NA 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: The paper details the WaTL and AWaTL algorithms (Algorithms 1 and 2 in Section 3), the data generation process for simulations (Section 5), and the preprocessing steps for the real-world NHANES data application and human mortality data application (Section 6, Appendix A), providing sufficient information to understand how the main experimental results were obtained.

Guidelines:

- The answer NA 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: [Yes]

Justification: The paper uses publicly available datasets, including NHANES (Section 6) and human mortality data (Appendix A), with URLs provided. The code used to generate all results is included in the submission as a zip folder, supporting reproducibility of the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.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://nips.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, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies experimental settings, including sample sizes, data generation procedures for simulations (Section 5), data preprocessing for real-world applications (Section 6), and how hyperparameters like λ were chosen (five-fold cross-validation in Section 5).

Guidelines:

- The answer NA 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: [No]

Justification: The paper reports average RMSPR over 50 repetitions for simulations (Section 5) and results from cross-validation for real data (Section 6), but error bars, confidence intervals, or formal statistical significance tests for these experimental results are not explicitly reported in the figures or text.

Guidelines:

- The answer NA 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
- 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: [No]

Justification: This paper introduces a new transfer learning method for regression with distributional outputs, emphasizing methodological and theoretical contributions rather than computational efficiency. The approach is not compute-intensive and was implemented using standard CPU resources, so details on compute infrastructure and execution time were not included.

Guidelines:

- The answer NA 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: We have reviewed the NeurIPS Code of Ethics and confirm that the research conforms to it. The work focuses on algorithmic development and utilizes publicly available, de-identified data for real-world applications (NHANES and human mortality datasets).

Guidelines:

- The answer NA 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: [NA]

Justification: The paper primarily focuses on the methodological and theoretical contributions of the proposed transfer learning framework. A dedicated discussion of potential positive and negative societal impacts is not included, though the real-world applications (Sections 6) in health and demography hint at positive uses.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The research introduces new regression algorithms and applies them to existing public datasets. It does not involve the release of new pretrained models or datasets that pose a high or direct risk for misuse in the manner of generative AI or sensitive scraped data.

Guidelines:

- The answer NA 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 credits the NHANES and human mortality datasets and provides the related URLs (Section 6).

Guidelines:

- The answer NA 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: [\[NA\]](#)

Justification: The paper introduces new algorithms (WaTL and AWaTL) which are described in detail (Algorithms 1 and 2, Section 3). No new datasets or software/models are being released as downloadable assets alongside the paper that would require separate documentation files.

Guidelines:

- The answer NA 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: [\[NA\]](#)

Justification: The research uses existing, publicly available, and de-identified datasets (NHANES, human mortality data from Section 6) involving human subjects. It does not involve new data collection through crowdsourcing or direct interaction with human subjects by the authors. Therefore, details of participant instructions or compensation for the original data collection are not applicable to this paper's contribution.

Guidelines:

- The answer NA 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: [NA]

Justification: The research involves the secondary analysis of publicly available and de-identified datasets (NHANES, human mortality data from Section 6). Such research typically does not require a new IRB approval for the current authors, as ethical oversight and risk disclosure were presumably handled by the original data collectors. The paper does not describe new risks to participants arising from this secondary analysis.

Guidelines:

- The answer NA 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: [NA]

Justification: The core methodology of this research focuses on transfer learning for regression models with distributional outputs in Wasserstein space and does not involve the use of Large Language Models as an important, original, or non-standard component.

Guidelines:

- The answer NA 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 (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Human Mortality Data

We further assess WaTL using the age-at-death distributions from 162 countries in 2015, compiled from the United Nations Databases⁴ and the UN World Population Prospects 2022⁵. The dataset provides country-level age-specific death counts, which we convert into smooth age-at-death densities using local linear smoothing; see Figure 3 (a) for an illustration. For this analysis, we define the 24 developed countries as the target site and the remaining 138 developing countries as the source site.

We evaluate the performance of WaTL against several key baselines on this dataset. These include models trained only on the target data (Only Target), only on the source data (Only Source), and on a naive pooling of both (Target + Source). Furthermore, for these baselines, we compare the performance of our underlying Fréchet regression framework against Wasserstein Regression [7], another state-of-the-art method for distributional data.

Table 1: Performance and Training Time with Varying Target Sample Sizes.

Number of Target Samples	RMSPR	Training Time (ms)
14	0.028	0.598
19	0.025	0.597
24	0.022	0.694

The comprehensive results are presented in Figure 3 (b). The proposed WaTL method achieves the lowest Root Mean Squared Prediction Risk (RMSPR) of 0.022, demonstrating a marked improvement over all alternatives. Notably, it significantly outperforms models trained solely on the 24 target samples (RMSPR of 0.027 for Fréchet Regression and 0.025 for Wasserstein Regression) and the naive data pooling approach (RMSPR of 0.033). This highlights that simply combining datasets is insufficient to overcome the domain shift between developing and developed countries, validating the necessity of our bias-correction mechanism.

To further assess WaTL’s robustness and practicality, especially in common scenarios with limited target data, we analyze its performance with a varying number of target samples. As shown in Table 1, the model’s predictive accuracy consistently improves as the target sample size increases from 14 to 24. It is also worth noting that the method maintains exceptional computational efficiency, with training times remaining under one millisecond. These findings confirm that WaTL is not only highly accurate but also a robust and efficient solution for real-world demographic studies where target data can be scarce.

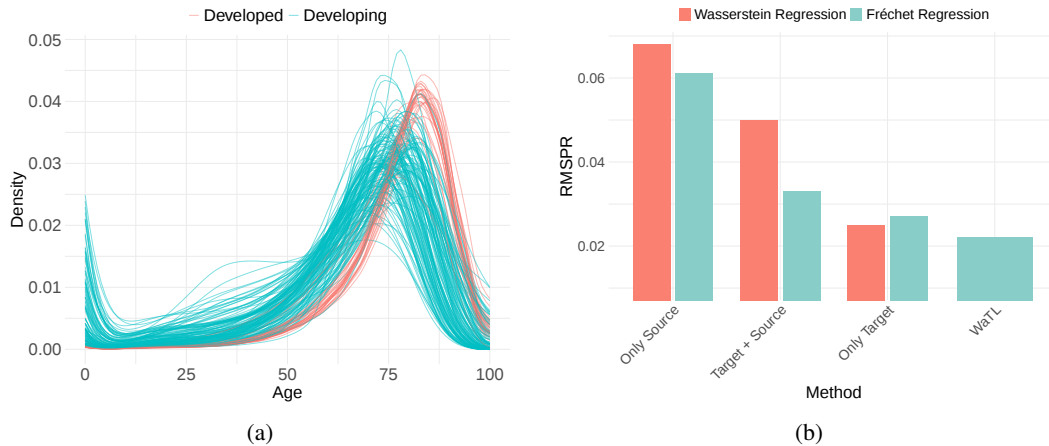


Figure 3: (a) Age-at-death densities of developed and developing countries; (b) Root mean squared prediction risk (RMSPR) of WaTL and Only Target methods for human mortality data.

⁴<https://data.un.org/>

⁵<https://population.un.org/wpp/Download>

B Additional Notations

For any $g_1, g_2 \in L^2(0, 1)$, the L^2 inner product is defined as

$$\langle g_1, g_2 \rangle_2 = \int_0^1 g_1(z)g_2(z)dz.$$

The total variation of a function g on the interval (a, b) is

$$V_a^b(g) = \lim_{n \rightarrow +\infty} \sup_{a < x_1 < \dots < x_n < b} \sum_{i=1}^n |g(x_i) - g(x_{i-1})|.$$

The space of bounded variation functions is given by

$$BV((0, 1), H) = \{g: (0, 1) \rightarrow (-H, H) | V_a^b(g) \leq H\},$$

which is equipped with the L^2 metric.

For a matrix Σ , let $\|\Sigma\|$ and $\|\Sigma\|_F$ denote its operator norm and Frobenius norm, respectively, and let $\gamma_i(\Sigma)$ be its i th smallest singular value.

We denote the characteristic function of a set $\mathcal{A}$ by $1_{\mathcal{A}}$ and use $|\mathcal{A}|$ to represent its cardinality. In a metric space (Ω, d) , the covering number $\mathcal{N}(K, d, \epsilon)$ represents the smallest number of closed balls of radius ϵ required to cover a subset $K \subset \Omega$. The notation $\sqcup$ is used to denote the disjoint union of sets.

For notational simplicity, we will omit (x) from $f(x)$ when the meaning is clear from the context.

C Proof

Lemma 2. Consider $U : \mathcal{W} \times (0, 1) \mapsto \mathbb{R}$, where $U(\nu, x) := F_{\nu}^{-1}(x)$ and $\mathcal{W} \times (0, 1)$ is endowed with the product σ -algebra generated by the Borel algebra on $\mathcal{W}$ and the Borel algebra on $(0, 1)$. The first Borel algebra $\mathcal{B}(\mathcal{W})$ is generated by open balls in $(\mathcal{W}, d_{\mathcal{W}})$ and the second $\mathcal{B}((0, 1))$ is generated by Euclidean open balls. Then U is measurable.

Lemma 2 enables a simplified expression for the conditional Fréchet mean of ν given $X = x$ by applying Fubini's theorem. Specifically, we have

$$\begin{aligned} m(x) &= \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu, \mu) | X = x\} \\ &= \arg \min_{\mu \in \mathcal{W}} E\{d_{L^2}^2(F_{\nu}^{-1}, F_{\mu}^{-1}) | X = x\} \\ &= \arg \min_{\mu \in \mathcal{W}} d_{L^2}^2(E\{F_{\nu}^{-1} | X = x\}, F_{\mu}^{-1}), \end{aligned}$$

where the last equality follows from Fubini's theorem, since $E\{F_{\nu}^{-1} | X = x\}$ is measurable. A similar argument applies to $m_G(x)$ and $\hat{m}_G(x)$.

C.1 Proof of Lemma 2

Proof. Without loss of generality, It suffices to prove $\{(\nu, x) | F_{\nu}^{-1}(x) > 0\}$ is measurable. Note that any quantile function is left continuous, hence $\{(\nu, x) | F_{\nu}^{-1}(x) > 0\} = \cup_{q \in \mathbb{Q} \cap (0, 1)} \{(\nu, q) | F_{\nu}^{-1}(q) > 0\}$.

Then without loss of generality, we suffice to prove $\{(\nu, 0.5) | F_{\nu}^{-1}(0.5) > 0\}$ is measurable. Since $\mathcal{W}$ is separable, we can select a dense countable subset K . We then define $A = \{\nu \in K | F_{\nu}^{-1}(0.5) > 0\}$. Assuming $A = \{\nu_i, i \in \mathbb{N}\}$, we have $\{\nu | F_{\nu}^{-1}(0.5) > 0\} = \{\nu | \lim_{i \rightarrow \infty} d_{\mathcal{W}}(\nu_i, \nu) = 0\} = \lim_{n \rightarrow \infty} \overline{\lim_{i \rightarrow \infty} B_{\frac{1}{n}}(\nu_i)}$ where $B_{\epsilon}(\nu)$ is the ball centered at ν with radius ϵ . Here we remark that both left continuous and monotone increasing are used in the first equation. Hence U is measurable. $\square$

C.2 Proof of Lemma 1

Proof. First, it is easy to check that $f^{(k)} \in BV((0, 1), H_0)$ for some $H_0 > 0$ since $f^{(k)} = E\{s_G^{(k)}(x)1_{\{s_G^{(k)}(x)>0\}}F_{\nu^{(k)}}^{-1}\} - E\{|s_G^{(k)}(x)|1_{\{s_G^{(k)}(x)<0\}}F_{\nu^{(k)}}^{-1}\}$ and $E|s_G^{(k)}(x)| < \infty$. By taking H_0 large enough, We also claim that $P(\hat{f}^{(k)} \in BV((0, 1), H_0)) \rightarrow 1$ since $\hat{f}^{(k)}$ also has a similar decomposition and

$$\begin{aligned} \frac{1}{n_k} \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x)| &\leq \frac{1}{n_k} \sum_{i=1}^{n_k} (|s_{iG}^{(k)}(x) - s_i^{(k)}(x)| + |s_i^{(k)}(x)|) \\ &\leq o_p(1) + H_1 \end{aligned}$$

for some $H_1 > 0$ with high probability, where $s_i^{(k)}(x) = 1 + (X_i^{(k)} - \theta_k)\Sigma_k^{-1}(x - \theta_k)$. The last inequality holds because of the result given by the proof of Theorem 2 in [26] that $\frac{1}{n_k} \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x) - s_i^{(k)}(x)| = o_p(1)$ and the assumption that $X^{(k)}$ is subgaussian.

To use Theorem 2 in [26], taking Ω to be $BV((0, 1), H_0)$, $M(g, x)$ to be $M^{(k)}(g, x) = E\{s_G^{(k)}(x)\|F_{\nu^{(k)}}^{-1} - g\|_2^2\}$ for any $g \in BV((0, 1), H_0)$, it suffices to check that the other two assumptions of the theorem:

$$\int_0^1 \sqrt{1 + \log \mathcal{N}(B_\delta(f^{(k)}) \cap BV((0, 1), H_0), d_{L^2}, \delta\epsilon)} d\epsilon = O(1)$$

as $\delta \rightarrow 0$ and

$$M^{(k)}(g, x) - M^{(k)}(f^{(k)}, x) \geq \|f^{(k)} - g\|_2^2.$$

Example 19.11 in [33] gives a bound of the covering number of the space $(BV((0, 1), H_0), d_{L^2})$,

$$\mathcal{N}(BV((0, 1), H_0), d_{L^2}, \epsilon) \leq e^{\frac{K}{\epsilon}}$$

for some $K > 0$ where K is independent of ϵ .

The rest is similar to the proof of Proposition 1 in [26]. For $g \in BV((0, 1), H_0)$, $B_\gamma(g)$ denotes the L^2 ball of radius γ centered at g . Let $\mathcal{C}_\epsilon(f^{(k)}) := \{g_u : u \in U\}$ such that $|U| = |\mathcal{N}(BV((0, 1), H_0) \cap B_1(f), d_{L^2}, \epsilon)| \leq e^{K\epsilon^{-1}}$ and the balls $B_\epsilon(g_u)$ covers $B_1(f) \cap BV((0, 1), H_0)$. For $\delta > 0$, we define $\tilde{g}_u = f^{(k)} + \delta(g_u - f^{(k)})$. Then the balls $B_{\delta\epsilon}(\tilde{g}_u)$ covers $B_\delta(f) \cap BV((0, 1), H_0)$. Hence

$$\int_0^1 \sqrt{1 + \log \mathcal{N}(B_\delta(m_G(x)) \cap BV((0, 1), H_0), d_{L^2}, \delta\epsilon)} d\epsilon \leq \int_0^1 \sqrt{1 + K\epsilon^{-1}} d\epsilon \leq 1 + 2\sqrt{K} < \infty.$$

For the last assumption, just note that

$$M^{(k)}(g, x) - M^{(k)}(f^{(k)}, x) = \|f^{(k)} - g\|_2^2.$$

Then according to Theorem 2 in [26], $\|\hat{f}^{(k)} - f^{(k)}\|_2 = O_p(n_k^{-1/2})$. $\square$

C.3 Proof of Theorem 1

To prove Theorem 1, we first establish a lemma that quantifies the bias introduced in Step 1 of Algorithm 1. This lemma is formulated for a general metric space, making it broadly applicable and potentially useful for extending transfer learning algorithms to other metric spaces.

Here are some notations, most of which are similar to our original setting. The only difference is that we use Y instead of ν to represent a random object in a general metric space (Ω, d) . We define $n_* = \min_{1 \leq k \leq K} n_k$, $n_{\mathcal{A}} = \sum_{i=1}^K n_i$, $\alpha_k = n_k/(n_0 + n_{\mathcal{A}})$,

$$m_1(x) = \arg \min_{\omega \in \Omega} M^1(\omega, x), \quad M^1(\omega, x) = \sum_{k=0}^K \alpha_k E\{s_G^{(k)}(x)d^2(\omega, Y_i^{(k)})\},$$

and

$$\hat{m}_1(x) = \arg \min_{\omega \in \Omega} M_n^1(\omega, x), \quad M_n^1(\omega, x) = \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K \sum_{i=1}^{n_k} s_{iG}^{(k)}(x)d^2(\omega, Y_i^{(k)}).$$

We impose the following mild condition on the metric space (Ω, d) , which is widely used in the literature on non-Euclidean data analysis [26]. This condition holds for various metric spaces commonly encountered in real-world applications, including the Wasserstein space, the space of networks, and the space of symmetric positive-definite matrices, among others.

Condition 4. (i) $m_1(x)$ uniquely exists and $\hat{m}_1(x)$ uniquely exists almost surely. Additionally for any $\gamma > 0$, $\inf_{d(m_1(x), \omega) > \gamma} M^1(\omega, x) > M^1(m_1(x), x)$.

(ii) Let $B_\delta(m_1(x)) \subset \Omega$ be the ball of radius δ centered at $m_1(x)$. Then

$$J(\delta) := \int_0^1 \sqrt{1 + \log \mathcal{N}(B_\delta(m_1(x)), d, \delta\epsilon)} d\epsilon = O(1)$$

as $\delta \rightarrow 0$.

(iii) There exist $\eta > 0$, $C > 0$ and $\beta > 1$, possibly depending on x , such that whenever $d(m_1(x), \omega) < \eta$, we have $M^1(\omega, x) - M_n^1(m_1(x), x) \geq Cd^\beta(\omega, m_1(x))$.

Lemma 3. Under Conditions 1, 2 and 4, $d^2(\hat{m}_1(x), m_1(x)) = O_p\left(\left(\frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_A} + (n_0 + n_A)^{-1/2}\right)^{\frac{1}{\beta-1}}\right)$.

C.3.1 Proof of Lemma 3

Proof. Denote $V_n(\omega, x) = M_n^1(\omega, x) - M^1(\omega, x)$, and $D_i^{(k)}(\omega, x) = d^2(Y_i^{(k)}, \omega) - d^2(Y_i^{(k)}, m_1(x))$ and recall $s_i^{(k)}(x) = 1 + (X_i^{(k)} - \theta_k) \Sigma_k^{-1}(x - \theta_k)$. Then

$$\begin{aligned} |V_n(\omega, x) - V_n(m_1(x), x)| &\leq \sum_{k=0}^K \alpha_k \left\{ \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (s_{iG}^{(k)}(x) - s_i^{(k)}(x)) D_i^{(k)}(\omega) \right| \right. \\ &\quad \left. + \left| \frac{1}{n_k} \sum_{i=1}^{n_k} s_i^{(k)} D_i^{(k)}(\omega) - E\{s_i^{(k)} D_i^{(k)}(\omega)\} \right| \right\}. \end{aligned} \quad (\text{A.1})$$

For any $\delta > 0$,

$$\sup_{d(\omega, m_1(x)) < \delta} \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (s_{iG}^{(k)}(x) - s_i^{(k)}(x)) D_i^{(k)}(\omega) \right| \leq \frac{2\text{diam}(\Omega)\delta}{n_k} \sum_{i=1}^{n_k} |W_0^{(k)}(x) + W_1^{(k)}(x)^\top X_i^{(k)}|,$$

where $W_0^{(k)}(x) := (\bar{X}^{(k)})^\top \Sigma_k^{-1}(x - \bar{X}^{(k)}) - \theta_k^\top \Sigma_k^{-1}(x - \theta_k)$ and $W_1^{(k)}(x) := \Sigma_k^{-1}(x - \theta_k) - \hat{\Sigma}_k(x - \bar{X}^{(k)})$. Then

$$s_{iG}^{(k)}(x) - s_i^{(k)}(x) = W_0^{(k)}(x) + (W_1^{(k)}(x))^\top X_i^{(k)}. \quad (\text{A.2})$$

We neglect the superscript for the sake of notation simplicity.

Denote $B_{1,M} := \{\|\frac{1}{n} \sum_{i=1}^n |X_i|\| \leq M\}$ where $|X_i|$ represents the element-wise absolute value of X_i . Let $B_{2,M} = \{\|\hat{\Sigma}^{-1}\| \leq M\}$, $B_M = B_{1,M} \cap B_{2,M}$ and $\hat{\Sigma}' = \frac{1}{n} \sum (X_i - \theta)(X_i - \theta)^\top$. In $B_{1,M}$,

$$\begin{aligned} \|\hat{\Sigma} - \hat{\Sigma}'\| &\leq \|\hat{\Sigma} - \hat{\Sigma}'\|_F \\ &= \left\| \frac{1}{n} \sum_{i=1}^n [(X_i - \bar{X})(X_i - \bar{X})^\top - (X_i - \theta)(X_i - \theta)^\top] \right\|_F \\ &\leq [2p^2 \|\bar{X}\|^2 - \theta^2\|_2^2 + M^2 \|\bar{X} - \theta\|_2^2]^{1/2} \\ &\leq 2p(M + |\theta|) \|\bar{X} - \theta\|_2. \end{aligned}$$

Using Theorem 6.5 in [37] leads to

$$P(\|\hat{\Sigma} - \Sigma\| > \epsilon) \leq 2e^{-C_1 \epsilon^2 n} + C_2 e^{-C_3(\frac{1}{2}\epsilon - \sqrt{\frac{C_4}{n}})^2 n},$$

for any $\epsilon \in (2\sqrt{\frac{C_4}{n}}, 1)$. Note that

$$\begin{aligned}\|\widehat{\Sigma}^{-1}\| &\leq \|\Sigma^{-1}\| + \|\widehat{\Sigma}^{-1} - \Sigma^{-1}\| \\ &\leq \|\Sigma^{-1}\| + \frac{|\gamma_1(\Sigma) - \gamma_1(\widehat{\Sigma})|}{\gamma_1(\Sigma)\gamma_1(\widehat{\Sigma})} \\ &\leq \|\Sigma^{-1}\| + \frac{\|\Sigma - \widehat{\Sigma}\|}{\gamma_1(\Sigma)(\gamma_1(\Sigma) - \|\Sigma - \widehat{\Sigma}\|)},\end{aligned}$$

where the third inequality is from the Weyl's inequality. Then there exists $M > 0$ such that

$$P(\|\widehat{\Sigma}^{-1}\| \geq M) \leq 2e^{-C_1 \frac{4}{R_3^2} n} + C_2 e^{(-C_3(\frac{2}{R_3} - \sqrt{\frac{C_4}{n}})n)}.$$

Consequently,

$$P(B_M^c) \leq 3e^{-\frac{4C_1}{R_3^2} n} + C_2 e^{(-C_3(\frac{2}{R_3} - \sqrt{\frac{C_4}{n}})n)}.$$

In B_M ,

$$\begin{aligned}W_0(x) &= |(\overline{X} - \theta)\Sigma^{-1}x - \theta\Sigma^{-1}(\theta - \overline{X}) - \overline{X}\Sigma^{-1}(\theta - \overline{X})| \\ &\leq |(\overline{X} - \theta)\Sigma^{-1}(|x| + |\theta| + M)|.\end{aligned}$$

Hence in $B'_M := \cap_{k=0}^K B_M^{(k)}$, we have

$$\begin{aligned}P\left(\sum_{k=0}^K \alpha_k |W_0(x)| > t\right) &\leq 2e^{-\frac{C_5 t^2}{\sum_{k=0}^K \alpha_k^2 \|W_0(x)\|_{\Psi_2}^2}} \\ &\leq 2e^{-\frac{C_6 t^2}{\sum_{k=0}^K \frac{n_k}{(n_0 + n_{\mathcal{A}})^2}}} \leq 2e^{-C_6 t^2 (n_0 + n_{\mathcal{A}})},\end{aligned}\tag{A.3}$$

for some constants $C_5, C_6 > 0$.

In the first inequality, we use general Hoeffding's inequality (Theorem 2.6.2 in [35]) and in the second inequality, we use Proposition 2.6.1 in [35].

In addition,

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n |(W_1(x))^T X_i| &\leq M[\|\Sigma^{-1} - \widehat{\Sigma}^{-1}\| \|x\| - \Sigma^{-1}(\theta - \overline{X}) + (\Sigma^{-1} - \widehat{\Sigma}^{-1})\overline{X}] \\ &\leq M(\|x\| + M)(\|\Sigma^{-1} - \widehat{\Sigma}^{-1}\|) + MR_3|\theta - \overline{X}| \\ &\leq M^2(\|x\| + M)R_3\|\Sigma - \widehat{\Sigma}\| + MR_3|\theta - \overline{X}|.\end{aligned}$$

Hence in B'_M

$$P\left(\sum_{k=0}^K \alpha_k \frac{1}{n_k} \sum_{i=1}^{n_k} |(X_i^{(k)})^T W_1^{(k)}(x)| > t\right) \leq 2e^{-\frac{C_8 t^2}{n_0 + n_{\mathcal{A}}}} + P\left(\sum_{k=0}^K \alpha_k C_9 \|\Sigma_k - \widehat{\Sigma}_k\| > t\right)\tag{A.4}$$

for some constant $C_8, C_9 > 0$. Using Markov inequality, the second term is bounded by

$$e^{-\lambda t} \prod_{k=0}^K E e^{\lambda \alpha_k C_9 \|\Sigma_k - \widehat{\Sigma}_k\|} \leq e^{-\lambda t} e^{4pK + \sum_{k=0}^K C_{10} \frac{\alpha_k^2 \lambda^2}{n_k}},$$

which is from Theorem 6.5 in [37] for any λ satisfying $\lambda \leq C_{11} n_*$ for some constants $C_{10}, C_{11} > 0$. We can choose $\lambda := \frac{t(n_0 + n_{\mathcal{A}})}{2C_{10}}$, which does satisfy the theorem's assumption if $t = O((n_0 + n_{\mathcal{A}})^{-1/2})$. Here we utilize Condition 2. Then

$$P\left(\sum_{k=0}^K \alpha_k C_9 \|\Sigma_k - \widehat{\Sigma}_k\| > t\right) \leq e^{-\frac{t^2(n_0 + n_{\mathcal{A}})}{4C_{12}} + 4dK}.\tag{A.5}$$

Note that $K = o(n_0 + n_{\mathcal{A}})$, hence

$$\sup_{d(\omega, m_1(x)) < \delta} \sum_{k=0}^K \alpha_k \left[\left| \frac{1}{n_k} \sum (s_{iG}^{(k)}(x) - s_i^{(k)}(x)) D_i^k(\omega) \right| \right] = O_p \left(\delta (n_0 + n_{\mathcal{A}})^{-1/2} \right),$$

since $P(B_M) \rightarrow 1$ due to Condition 2.

To bound the second term of (A.1), following the proof of Theorem 2 in [26] for each k , we define the functions $g_{\omega}^{(k)} : \mathbb{R}^p \times \Omega \mapsto \mathbb{R}$ as

$$g_{\omega}^{(k)}(z, y) = [1 + (z - \theta_k)^{\top} \Sigma_k^{-1} (x - \theta_k)] d^2(y, \omega)$$

and the function class

$$\mathcal{M}_{\delta}^{(k)} := \{g_{\omega}^{(k)} - g_{m_1(x)}^{(k)} : d(\omega, m_1(x)) < \delta\}.$$

We have

$$E \left\{ \sup_{d(\omega, m_1(x)) < \delta} \left| \frac{1}{n_k} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) D_{(i)}^k(g) - E\{s_i^{(k)} D_i^{(k)}(g)\} \right| \right\} \leq J(E[(G_{\delta}^{(k)}(x))^2]^{\frac{1}{2}} \sqrt{n_k},$$

where $G_{\delta}^{(k)}(z) := 2\text{diam}(\Omega)\delta[1 + (z - \theta_k)^{\top} \Sigma_k^{-1} (x - \theta_k)]$ is the envelop function.

Note that

$$E(G_{\delta}^{(k)}(X)^2) \leq C_{14}\delta^2$$

for some constant $C_{14} > 0$, which does not depend on k . Hence

$$E \left\{ \sup_{d(\omega, m_1(x)) < \delta} \left| \sum_{k=0}^K \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) D_i^{(k)}(g) - E\{s_i^{(k)} D_i^{(k)}(g)\} \right| \right\} = O \left(\delta \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right). \quad (\text{A.6})$$

Define

$$D_R := \left\{ \sup_{d(\omega, m_1(x)) < \delta} \sum_{k=0}^K \alpha_k \left| \frac{1}{n_k} \sum_{i=1}^{n_k} (s_{iG}^{(k)}(x) - s_i^{(k)}(x)) D_i^{(k)}(\omega) \right| \leq R\delta(n_0 + n_{\mathcal{A}})^{-1/2} \right\}.$$

We have

$$E\{1_{D_R} \sup_{d(\omega, m_1(x)) < \delta} |V_n(w) - V_n(m_1(x))|\} \leq aR\delta \left(\sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_o + n_{\mathcal{A}})^{-1/2} \right),$$

for some constant $a > 0$.

Next we show

$$|M^1(w, x) - M_n^1(w, x)| = o_p(1),$$

and for all $\omega \in \Omega$ and for any $\kappa > 0, \varphi > 0$, there exists $\delta > 0$ such that

$$\limsup_n P \left(\sup_{d(\omega_1, \omega_2) < \delta} |M_n^1(\omega_1, x) - M_n^1(\omega_2, x)| > \kappa \right) \leq \varphi. \quad (\text{A.7})$$

To prove the first assertion, we denote

$$\tilde{M}_n^1(\omega, x) = \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K \sum_{i=1}^{n_k} s_i^{(k)}(x) d^2(\omega, Y_i^{(k)}).$$

Then $E\tilde{M}_n^1(\omega, x) = M_n^1(\omega, x)$ and $\text{Var}(\tilde{M}_n^1(\omega, x)) \leq 2 \sum_{k=0}^K C' \frac{n_k}{(n_0 + n_{\mathcal{A}})^2}$ for some constant $C' > 0$. Hence $\tilde{M}_n^1(\omega, x) - M_n^1(\omega, x) = o_p(1)$.

Besides,

$$\begin{aligned} M_n^1(\omega) - \tilde{M}_n^1(\omega) &= \sum_{k=0}^K \alpha_k \frac{W_0^{(k)}}{n_k} \sum_{i=1}^{n_k} d^2(Y_i^{(k)}, \omega) + \sum_{k=0}^K \alpha_k \frac{(W_1^{(k)})^{\top}}{n_k} \sum_{i=1}^{n_k} X_i^{(k)} d^2(Y_i^{(k)}, \omega) \\ &= o_p(1). \end{aligned}$$

The last equation is true since $\sum_{k=0}^K \alpha_k \frac{1}{n_k} \sum_{i=1}^{n_k} |(W_1(x))^T X_i| = o_p(1)$ and $\sum_{k=0}^K \alpha_k |W_0(x)| = o_p(1)$.

To prove (A.7), note that for any $\gamma_1, \gamma_2 \in \Omega$,

$$\begin{aligned} |M_n^1(\gamma_1, x) - M_n^1(\gamma_2, x)| &\leq 2\text{diam}(\Omega)d(\gamma_1, \gamma_2) \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K \sum_{i=1}^{n_k} |s_i^{(k)} + W_0^{(k)} + (W_1^{(k)})^T X_i| \\ &= O_p(d(\gamma_1, \gamma_2)). \end{aligned}$$

The last equation is true since $\sum_{k=0}^K \alpha_k \frac{1}{n_k} \sum_{i=1}^{n_k} |(W_1(x))^T X_i| = o_p(1)$ and $\sum_{k=0}^K \alpha_k |W_0(x)| = o_p(1)$, and we can prove

$$\sum_{k=0}^K \alpha_k \frac{1}{n_k} \sum_{i=1}^{n_k} |s_i^{(k)}| = 1 + o_p(1) \quad (\text{A.8})$$

in a similar way of (A.3). It follows that $d(\hat{m}_1(x), m_1(x)) = o_p(1)$.

Set $r_n = \left(\sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-1/2} \right)^{\frac{\beta}{2(\beta-1)}}$ and

$$S_{j,n}(x) = \{\omega : 2^{j-1} < r_n d(\omega, m_1(x))^{\frac{\beta}{2}} \leq 2^j\}.$$

Choose $\eta > 0$ to satisfy (iii) in Condition 4 and also small enough that (ii) in Condition 4 holds for all $\delta < \eta$ and set $\tilde{\eta} := \eta^{\frac{\beta}{2}}$. For any integer L_2 ,

$$\begin{aligned} P\left(r_n d^{\beta/2}(\hat{m}_1(x), m_1(x)) > 2^{L_2}\right) &\leq P(D_R^c) + P\left(2d(\hat{m}_1(x), m_1(x)) \geq \eta\right) \\ &+ \sum_{\substack{j \geq L_2 \\ 2^j \leq r_n \tilde{\eta}}} P\left(\left\{ \sup_{\omega \in S_{j,n}} |V_n(\omega) - V_n(m_1(x))| \geq C \frac{2^{2(j-1)}}{r_n^2} \right\} \cap D_R\right). \end{aligned}$$

The second term converges to 0 since $d(\hat{m}_1(x), m_1(x)) = o_p(1)$. For each j in the sum in the third term, we have $d(\omega, m_1(x)) \leq \left(\frac{2^j}{r_n}\right)^{\frac{2}{\beta}} \leq \eta$, so the sum is bounded by

$$4aC^{-1} \sum_{\substack{j \geq L_2 \\ 2^j \leq r_n \tilde{\eta}}} \frac{2^{\frac{2j(1-\beta)}{\beta}}}{r_n^{\frac{2(1-\beta)}{\beta}} \left(\sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-\frac{1}{2}} \right)} \leq 4aC^{-1} \sum_{j \geq L_2} \left(\frac{1}{4^{\frac{\beta-1}{\beta}}} \right)^j.$$

Choose L_2 large enough, this probability can be small enough.

Hence we have

$$d^2(\hat{m}_1(x), m_1(x)) = O_p\left(\frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-1/2}\right)^{\frac{1}{\beta-1}}.$$

□

C.3.2 Proof of Theorem 1 Given Lemma 3

Proof. The proof is similar to that of Lemma 1. First, it is easy to check that $f \in BV((0, 1), H_2)$ for some large $H_2 > 0$ since

$$f = E \left\{ \sum_{k=0}^K \alpha_k s_G^{(k)}(x) 1_{\{\sum_{k=0}^K \alpha_k s_G^{(k)}(x) > 0\}} F_{\nu^{(0)}}^{-1} \right\} - E \left\{ \left| \sum_{k=0}^K \alpha_k s_G^{(k)}(x) \right| 1_{\{\sum_{k=0}^K \alpha_k s_G^{(k)}(x) < 0\}} F_{\nu^{(0)}}^{-1} \right\}$$

and $E|\sum_{k=0}^K \alpha_k s_G^{(k)}(x)| < \infty$. Taking H_2 large enough, we also claim that $P(\hat{f} \in BV((0, 1), H_2)) \rightarrow 1$ since $\hat{f}_0$ also has a similar decomposition and

$$\begin{aligned} \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x)| &\leq \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{i=1}^{n_k} (|s_{iG}^{(k)}(x) - s_i^{(k)}(x)| + |s_i^{(k)}(x)|) \\ &\leq o_p(1) + H_3, \end{aligned}$$

for some $H_3 > 0$ with high probability. The last inequality follows from (A.2), (A.8), (A.5) and (A.4). Thus, (i) in Condition 4 holds.

(ii) in Condition 4 follows from the same arguments used in the proof of Lemma 1. It remains to verify (iii) in Condition 4, which holds since

$$M^1(g, x) - M^1(f, x) = \|f - g\|_2^2.$$

Hence the result follows by using Lemma 3. $\square$

C.4 Proof of Theorem 2

Proof. Recall that $f^{(0)}(x) = E\{s_G^{(0)}(x)F_{\nu^{(0)}}^{-1}\}$. Using the definition of $\hat{f}_0$,

$$\begin{aligned} \|\hat{f}_0 - f^{(0)}\|_2^2 &= \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) (\|\hat{f}_0 - F_{\nu_i^{(0)}}^{-1}\|_2^2 - \|f^{(0)} - F_{\nu_i^{(0)}}^{-1}\|_2^2) \\ &\quad + \frac{2}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) \langle F_{\nu_i^{(0)}}^{-1} - f^{(0)}, \hat{f}_0 - f^{(0)} \rangle_2 \\ &\leq -\lambda \|\hat{f}_0 - \hat{f}\|_2 + \lambda \|f^{(0)} - \hat{f}\|_2 \\ &\quad + \frac{2}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) \langle F_{\nu_i^{(0)}}^{-1} - f^{(0)}, \hat{f}_0 - f^{(0)} \rangle_2 \\ &\leq -\lambda \|\hat{f}_0 - \hat{f}\|_2 + \lambda \|f^{(0)} - \hat{f}\|_2 \\ &\quad + 2 \|\hat{f}_0 - f^{(0)}\|_2 \left\| \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) F_{\nu_i^{(0)}}^{-1} - f^{(0)} \right\|_2. \end{aligned}$$

We define the event $E_n := \{\|n_0^{-1} \sum_{i=0}^{n_0} s_{iG}^{(0)}(x) F_{\nu_i^{(0)}}^{-1} - f^{(0)}\|_2 \leq \lambda/2\}$. According to Lemma 1, we have $P(E_n) \rightarrow 1$ since $\lambda \asymp n_0^{-1/2+\epsilon}$.

Under E_n for n large enough,

$$\|\hat{f}_0 - f^{(0)}\|_2^2 \leq 2\lambda \|f^{(0)} - \hat{f}\|_2 \leq 2\lambda (\|f^{(0)} - \hat{f}\|_2 + \|\hat{f} - f\|_2)$$

holds. Hence we have

$$d_{\mathcal{W}}^2(\hat{m}_G^{(0)}(x), m_G^{(0)}(x)) \leq \|\hat{f}_0 - f^{(0)}\|_2^2 = O_p\left(n_0^{-1/2+\epsilon} (\|\hat{f} - f\|_2 + \psi)\right).$$

Then using Theorem 1, it follows that

$$d_{\mathcal{W}}^2(\hat{m}_G^{(0)}(x), m_G^{(0)}(x)) = O_p\left(n_0^{-1/2+\epsilon} \left(\psi + \frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-1/2}\right)\right).$$

$\square$

C.5 Proof of Theorem 3

Proof. We first prove $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$. Note that

$$\begin{aligned} P(\hat{\mathcal{A}} = \mathcal{A}) &\geq P(\max_{k \in \mathcal{A}} \hat{\psi}_k < \min_{k \in \mathcal{A}^C} \hat{\psi}_k) \\ &\geq 1 - P(\max_{1 \leq k \leq K} |\hat{\psi}_k - \psi_k| > |\max_{k \in \mathcal{A}} \psi_k - \min_{k \in \mathcal{A}^C} \psi_k|) \\ &\geq 1 - K \max_{1 \leq k' \leq K} P(|\hat{\psi}_{k'} - \psi_{k'}| > |\max_{k \in \mathcal{A}} \psi_k - \min_{k \in \mathcal{A}^C} \psi_k|). \end{aligned}$$

According to Lemma 1, $|\psi_k - \hat{\psi}_k| = O_p(n_k^{-1/2} + n_0^{-1/2})$. According to Condition 3,

$$\frac{|\max_{k \in \mathcal{A}} \psi_k - \min_{k \in \mathcal{A}^C} \psi_k|}{n_*^{-1/2} + n_0^{-1/2}} \rightarrow \infty.$$

Hence

$$P(\widehat{\mathcal{A}} = \mathcal{A}) = 1 - o(1).$$

Then using Theorem 2 where we substitute $\mathcal{A}$ for $\{1, \dots, K\}$, we have

$$d_{\mathcal{W}}^2(\widehat{m}_G^{(0)}(x), m_G^{(0)}(x)) = O_p\left(n_0^{-1/2+\epsilon}\left(\max_{k \in \mathcal{A}} \psi_k + \frac{\sum_{k \in \mathcal{A} \cup \{0\}} \sqrt{n_k}}{\sum_{k \in \mathcal{A} \cup \{0\}} n_k} + \left(\sum_{k \in \mathcal{A} \cup \{0\}} n_k\right)^{-1/2}\right)\right).$$

□

D Transfer Learning for Local Fréchet Regression

For simplicity, we consider a scalar predictor $X^{(k)} \in \mathbb{R}$ for $k = 0, \dots, K$. For predictors in $\mathbb{R}^p$, the explicit form of the weight function can be found in Section 2.3 of [17]. Under the same source-target data setting, we define

$$\begin{aligned} m^{(k)}(x) &= \arg \min_{\mu \in \mathcal{W}} E\{d_{\mathcal{W}}^2(\nu^{(k)}, \mu) | X^{(k)} = x\}, \\ m_{L,h}^{(k)}(x) &= \arg \min_{\mu \in \mathcal{W}} E\{s_L^{(k)}(x, h) d_{\mathcal{W}}^2(\nu^{(k)}, \mu)\}, \\ \widehat{m}_{L,h}^{(k)}(x) &= \arg \min_{\mu \in \mathcal{W}} \frac{1}{n_k} \sum_{i=1}^{n_k} s_{iL}^{(k)}(x, h) d_{\mathcal{W}}^2(\nu_i^{(k)}, \mu)\}, \end{aligned}$$

where $s_L^{(k)}(x)$ and $s_{iL}^{(k)}(x)$ represents the population and sample weight functions of local Fréchet regression for the k th source.

For notational simplicity, define

$$\begin{aligned} f_{\oplus}^{(k)}(x) &= E\{F_{\nu^{(k)}}^{-1} | X^{(k)} = x\}, \\ f_h^{(k)}(x) &= E\{s_L^{(k)}(x, h) F_{\nu^{(k)}}^{-1}\}, \\ f_h(x) &= \sum_{k=0}^K \alpha_k f_h^{(k)}(x). \end{aligned}$$

Similar to Section 3, we introduce the Local Wasserstein Transfer Learning (LWaTL) algorithm in Algorithm 3, assuming that all sources are informative. When the informative set is unknown, we incorporate an additional step to identify the informative set, as outlined in Algorithm 4. Theoretical guarantees for these algorithms are provided in Theorem 4 and Theorem 5.

We impose the following condition, where the first two parts correspond to the kernel and distributional assumptions that are standard in local linear regression.

Condition 5. (i) The kernel K is a probability density function, symmetric around zero. Furthermore, defining $K_{sj} = \int_{\mathbb{R}} K^s(u) u^j du$, $|K_{14}|$ and $|K_{26}|$ are both finite.

(ii) The marginal density $q^{(k)}$ of $X^{(k)}$ and the conditional density of $X^{(k)}$ given $\nu^{(k)}$, $g_{\nu^{(k)}}^{(k)}$, exist and are twice continuously differentiable. Besides, it satisfies that

$$\sup_{0 \leq k \leq K} \sup_{z, \nu^{(k)}} \max\{(q^{(k)})''(z), (g_{\nu^{(k)}}^{(k)})''(z)\} < H_3$$

for some $H_3 > 0$.

(iii) $h \rightarrow 0$ and $n_0 h \rightarrow \infty$.

(iv) $\frac{\sum_{k=1}^K n_k^{-1}}{h} = o(1)$.

(v) $\nu^{(k)}$ have a common bounded domain.

(vi) $u_j^{(k)} := E(K_h(X_i^{(k)} - x)(X_i^{(k)} - x)^j)$, $j = 0, 1, 2$ and $\sigma_0^{(k)} := u_0^{(k)} u_2^{(k)} - (u_1^{(k)})^2 > 0$ are bounded.

Algorithm 3 Local Wasserstein Transfer Learning (LWaTL)

Input: Target and source data $\{(x_i^{(0)}, \nu_i^{(0)})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \nu_i^{(k)})\}_{i=1}^{n_k})$, regularization parameter λ , bandwidth h and query point $x \in \mathbb{R}$.

Output: Target estimator $\hat{m}_{L,h}^{(0)}(x)$.

1: Weighted auxiliary estimator

$$\hat{f}_h(x) = \frac{1}{n_0 + n_A} \sum_{k=0}^K n_k \hat{f}_h^{(k)}(x),$$

where $\hat{f}_h^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iL}^{(k)}(x, h) F_{\nu_i^{(k)}}^{-1}$ and $n_A = \sum_{k=1}^K n_k$.

2: Bias correction using target data

$$\hat{f}_{0h}(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iL}^{(0)}(x, h) \|F_{\nu_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}_h(x)\|_2.$$

3: Projection to Wasserstein space

$$\hat{m}_{L,h}^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_{0h}(x)\|_2.$$

The following theorem establishes the rate of convergence for the LWaTL algorithm.

Theorem 4. Assume Condition 5 holds and the regularization parameter satisfies $\lambda \asymp n_0^{-1/2} h^{-1/2-\epsilon}$ for some $\epsilon > 0$. Then, for the LWaTL algorithm and a fixed $x \in \mathbb{R}^p$, we have

$$d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) = O_p \left(h^4 + n_0^{-1/2} h^{-1/2-\epsilon} \left[\psi_L + h^2 + h^{-1/2} \left(\left(\sum_{k=0}^K n_k^{-1} \right)^{1/2} + \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_A} \right) \right] \right),$$

where $\psi_L = \max_{1 \leq k \leq K} \|f_\oplus^{(0)}(x) - f_\oplus^{(k)}(x)\|$.

Proof. Note that

$$\begin{aligned} d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) &\leq 2\{d_{\mathcal{W}}^2(m_{L,h}^{(0)}(x), m^{(0)}(x)) + d_{\mathcal{W}}^2(m_{L,h}^{(0)}(x), \hat{m}_{L,h}^{(0)}(x))\} \\ &\leq 2\{d_{\mathcal{W}}^2(m_{L,h}^{(0)}(x), m^{(0)}(x)) + \|f_h^{(0)}(x) - \hat{f}_{0h}(x)\|_2^2\}. \end{aligned}$$

For the first term, We could use Theorem 3 of [26] since we can check all its assumptions easily, which is similar to the proof of Lemma 1. Here we should consider the metric space $BV((0, 1), H_4)$ for some $H_4 > 0$ endowed with the probability measure naturally induced by the canonical embedding of the Wasserstein space.

For the second term, first note that using Theorem 4 of [26], we have $\|f_h^{(0)}(x) - \hat{f}_h^{(0)}(x)\|_2 = O_p((n_0 h)^{-1/2})$. Then similar to the proof of Theorem 2, we have

$$\begin{aligned} \|f_h^{(0)}(x) - \hat{f}_{0h}(x)\|_2^2 &\leq -\lambda \|\hat{f}_{0h}(x) - \hat{f}_h(x)\|_2 + \lambda \|f_h^{(0)}(x) - \hat{f}_h(x)\|_2 \\ &\quad + 2\|\hat{f}_h^{(0)}(x) - f_h^{(0)}(x)\|_2 \|\hat{f}_{0h}(x) - f_h^{(0)}(x)\|_2. \end{aligned}$$

Hence in $E_n := \{\|f_h^{(0)}(x) - \hat{f}_h^{(0)}(x)\|_2 \leq \frac{1}{2}\lambda\}$,

$$\|f_h^{(0)}(x) - \hat{f}_{0h}(x)\|_2^2 \leq 2\lambda \|f_h^{(0)}(x) - \hat{f}_h(x)\|_2 \leq 2\lambda(\psi_h + \|f_h(x) - \hat{f}_h(x)\|_2), \quad (\text{A.9})$$

where $\psi_h = \max_{1 \leq k \leq K} \|f_h^{(k)} - f_h^{(0)}\|_2$. For the first term of (A.9), using Theorem 3 of [26] and (i), (ii) in Condition 5, we have

$$\psi_h = \psi_L + O(h^2).$$

For the second term of (A.9), we could follow the proof of Lemma 3 to study the asymptotic rate of convergence. The difference is that we do not assume covariates are sub-Gaussian but

Algorithm 4 Adaptive Local Wasserstein Transfer Learning (ALWaTL)

Input: Target and source data $\{(x_i^{(0)}, \nu_i^{(0)})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \nu_i^{(k)})\}_{i=1}^{n_k})$, regularization parameter λ , bandwidth h , number of informative sources L , and query point $x \in \mathbb{R}$.

Output: Target estimator $\hat{m}_{L,h}^{(0)}(x)$.

- 1: Compute discrepancy scores. For each source dataset $k = 1, \dots, K$, compute the empirical discrepancy

$$\hat{\psi}_{k,h} = \|\hat{f}_h^{(0)}(x) - \hat{f}_h^{(k)}(x)\|_2,$$

where $\hat{f}_h^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iL}^{(k)}(x, h) F_{\nu_i^{(k)}}^{-1}$. Construct the adaptive informative set by selecting the L smallest discrepancy scores

$$\hat{\mathcal{A}} = \{1 \leq k \leq K : \hat{\psi}_{k,h} \text{ is among the smallest } L \text{ values}\}.$$

- 2: Weighted auxiliary estimator

$$\hat{f}_h(x) = \frac{1}{\sum_{k \in \hat{\mathcal{A}} \cup \{0\}} n_k} \sum_{k \in \hat{\mathcal{A}} \cup \{0\}} n_k \hat{f}_h^{(k)}(x).$$

- 3: Bias correction using target data

$$\hat{f}_{0h}(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iL}^{(0)}(x, h) \|F_{\nu_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}_h(x)\|_2.$$

- 4: Projection to Wasserstein space

$$\hat{m}_{L,h}^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_{0h}(x)\|_2.$$

we have upper bounds for moments related to covariates and the kernel. In detail, we define $\tilde{s}_{iL}^{(k)}(x, h) := K_h(X_i^{(k)} - x) \frac{u_0^{(k)} - u_1^{(k)}(X_i^{(k)} - x)}{(\sigma_0^{(k)})^2}$ and $D_i^{(k)}(g, x) := \|F_{\nu_i^{(k)}}^{-1} - g\|_2^2 - \|F_{\nu_i^{(k)}}^{-1} - \hat{f}_h^{(k)}(x)\|_2^2$. Then similar to (A.6) and the proof of Theorem 4 in [26] we have for small $\delta > 0$,

$$\begin{aligned} & E \left\{ \sup_{d_{L^2}(g, \hat{f}_h^{(k)}) < \delta} \left| \frac{1}{n_0 + n_{\mathcal{A}}} \sum_{k=0}^K \sum_{i=1}^{n_k} \tilde{s}_{iL}^{(k)}(x, h) D_i^{(k)}(g, x) - \sum_{k=0}^K \alpha_k E \{ \tilde{s}_{iL}^{(k)}(x, h) D_i^{(k)}(g, x) \} \right| \right\} \\ &= O \left(\delta h^{-1/2} \frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right). \end{aligned}$$

Another term we need to bound is $\sum_{k=0}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h) - \tilde{s}_{iL}^{(k)}(x, h)|$. Note that we can define $B_M := \cap_{k=0}^K (\{|\hat{\sigma}_0^{(k)}|^2 > \frac{1}{2}(\sigma_0^{(k)})^2\} \cap \cap_{j=0}^2 (\{\frac{1}{n_k} \sum_{i=1}^{n_k} |K_h(X_i^{(k)} - x)(X_i^{(k)} - x)^j| < h^j M\}))$ for some large $M > 0$. Then it is easy to check that

$$P((B_M)^c) \leq \sum_{k=0}^K \frac{C_{A1}}{n_k},$$

for some $C_{A1} > 0$. In B_M , observe that

$$|s_{iL}^{(k)}(x, h) - \tilde{s}_{iL}^{(k)}(x, h)| \leq |W_{0n}^{(k)} K_h(X_i^{(k)} - x)| + |W_{1n}^{(k)} K_h(X_i^{(k)} - x)(X_i^{(k)} - x)|,$$

where

$$W_{0n}^{(k)} := \frac{\hat{u}_2^{(k)}}{(\hat{\sigma}_0^{(k)})^2} - \frac{u_2^{(k)}}{(\sigma_0^{(k)})^2}, \quad W_{1n}^{(k)} := \frac{\hat{u}_1^{(k)}}{(\hat{\sigma}_0^{(k)})^2} - \frac{u_1^{(k)}}{(\sigma_0^{(k)})^2}.$$

Note that in B_M , it is easy to check

$$|W_{0n}^{(k)}| \leq C_{A2} \left(\sum_{j=0}^2 h^{-j} |u_j^{(k)} - \hat{u}_j^{(k)}| \right)$$

and

$$|W_{1n}^{(k)}| \leq C_{A2} \left(\sum_{j=0}^2 h^{-1-j} |u_j^{(k)} - \hat{u}_j^{(k)}| \right)$$

for some constants $C_{A2} > 0$. Hence in B_M ,

$$\begin{aligned} & P \left(\sum_{k=0}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h) - \hat{s}_{iL}^{(k)}(x, h)| > t \right) \\ & \leq P \left(C_{A3} \sum_{k=0}^K \alpha_k (|W_{0n}| + h|W_{1n}|) > t \right) \\ & \leq \frac{1}{t^2 h} C_{A4} \sum_{k=0}^K \frac{1}{n_k}. \end{aligned}$$

Hence $\sum_{k=0}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h) - \hat{s}_{iL}^{(k)}(x, h)| = O_p(h^{-1/2} \sum_{k=0}^K n_k^{-1})$. In addition, we can check that $\hat{f}_h(x) \in BV((0, 1), H_5)$ exists almost surely for some $H_5 > 0$ by showing $\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h)| = O_p(1)$. Also utilizing the same technique in the proof of Lemma 3, It suffices to show that $\|f_h(x) - \hat{f}_h(x)\|_2 = o_p(1)$. It is just a simple generalization of Lemma 2 in [26]. Eventually, we have

$$\|f_h^{(0)}(x) - \hat{f}_{0h}(x)\|_2 = O_p \left(h^{-1/2} \left(\left(\sum_{k=0}^K n_k^{-1} \right)^{1/2} + \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right) \right)$$

and

$$d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) = O_p \left(h^4 + n_0^{-1/2} h^{-1/2-\epsilon} \left[\psi_L + h^2 + h^{-1/2} \left(\left(\sum_{k=0}^K n_k^{-1} \right)^{1/2} + \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right) \right] \right).$$

□

Remark 3. Theorem 4 can be extended to general metric spaces and serves as a parallel version of Lemma 3. Besides, if n_* is much larger than n_0 and $\psi_L = O(n_0^{-k})$ with $k > \frac{6}{15-10\epsilon}$, then among bandwidth sequences $h = n_0^{-r}$, the optimal sequence is achieved at $r^* = \frac{k}{2}$, leading to the convergence rate

$$d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) = O_p(n_0^{-\frac{3k-2\epsilon k+2}{4}}),$$

which surpasses the convergence rate of local Fréchet regression $O_p(n_0^{-4/5})$ [26].

There is also a parallel version of Theorem 3, built upon the following condition.

Condition 6. Suppose the regularization parameter satisfies $\lambda \asymp n_0^{-1/2} h^{-1/2-\epsilon}$ for some $\epsilon > 0$ and for some $\epsilon' > \epsilon$, There exists a non-empty subset $\mathcal{A} \subset \{1, \dots, K\}$ such that

$$\frac{h^2 + (n_* h)^{-1/2} + (n_0 h)^{-1/2}}{\min_{k \in \mathcal{A}^C} \psi_{k,L}} = o(1), \quad \frac{\max_{k \in \mathcal{A}} \psi_{k,L}}{h^2 + (n_* h)^{-1/2} + n_0^{-1/2} h^{-1/2+\epsilon'}} = O(1),$$

where $\mathcal{A}^C = \{1, \dots, K\} \setminus \mathcal{A}$, $n_* = \min_{1 \leq k \leq K} n_k$ and $\psi_{k,L} = \|f_{\oplus}^{(0)}(x) - f_{\oplus}^{(k)}(x)\|_2$.

Theorem 5. Assume Conditions 5 and 6 hold. Then for the ALWaTL algorithm we have

$$\begin{aligned} & d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m(x)) \\ & = O_p \left(h^4 + n_0^{-1/2} h^{-1/2-\epsilon} \left[\max_{k \in \mathcal{A}} \psi_{k,L} + h^2 \right. \right. \\ & \quad \left. \left. + h^{-1/2} \left(\left(\sum_{k \in \mathcal{A} \cup \{0\}} n_k^{-1} \right)^{1/2} + \sum_{k \in \mathcal{A} \cup \{0\}} \frac{\sqrt{n_k}}{\sum_{k \in \mathcal{A} \cup \{0\}} n_k} \right) \right] \right). \end{aligned}$$

Proof. We first prove $P(\hat{\mathcal{A}} = \mathcal{A}) \rightarrow 1$. Note that

$$\begin{aligned} P(\hat{\mathcal{A}} = \mathcal{A}) &\geq P(\max_{k \in \mathcal{A}} \hat{\psi}_{k,h} < \min_{k \in \mathcal{A}^C} \hat{\psi}_{k,h}) \\ &\geq 1 - P(\max_{1 \leq k \leq K} (|\hat{\psi}_{k,h} - \psi_{k,h}| + |\psi_{k,h} - \psi_{k,L}|) > |\max_{k \in \mathcal{A}} \psi_{k,L} - \min_{k \in \mathcal{A}^C} \psi_{k,L}|) \\ &\geq 1 - K \max_{1 \leq k' \leq K} P(|\hat{\psi}_{k',h} - \psi_{k'}| + |\psi_{k',h} - \psi_{k',L}|) > |\max_{k \in \mathcal{A}} \psi_{k,L} - \min_{k \in \mathcal{A}^C} \psi_{k,L}|. \end{aligned}$$

According to Theorem 3 of [26], $|\psi_{k,h} - \psi_{k,L}| = O(h^2)$. Utilizing Theorem 4 of [26], $|\psi_{k,h} - \hat{\psi}_{k,h}| = O_p((n_k h)^{-1/2} + (n_0 h)^{-1/2})$. According to Condition 6,

$$\frac{|\max_{k \in \mathcal{A}} \psi_{k,L} - \min_{k \in \mathcal{A}^C} \psi_{k,L}|}{h^2 + (n_* h)^{-1/2} + (n_0 h)^{-1/2}} \rightarrow \infty.$$

Hence

$$P(\hat{\mathcal{A}} = \mathcal{A}) = 1 - o(1).$$

Then utilizing Theorem 4 where we substitute $\mathcal{A}$ for $\{1, \dots, K\}$, we have

$$\begin{aligned} d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) &= O_p \left(h^4 + n_0^{-1/2} h^{-1/2-\epsilon} \left[\max_{k \in \mathcal{A}} \psi_{k,L} + h^2 \right. \right. \\ &\quad \left. \left. + h^{-1/2} \left(\left(\sum_{k \in \mathcal{A} \cup \{0\}} n_k^{-1} \right)^{1/2} + \sum_{k \in \mathcal{A} \cup \{0\}} \frac{\sqrt{n_k}}{\sum_{k \in \mathcal{A} \cup \{0\}} n_k} \right) \right] \right). \end{aligned}$$

□

Remark 4. If n_* is much larger than n_0 , then the simplified rate is $O_p(h^4 + n_0^{-1/2} h^{3/2-\epsilon} + n_0^{-1} h^{-1+\epsilon'-\epsilon})$. Among bandwidth sequences $h = n_0^r$, the optimal choice is achieved at $r^* = -\frac{1}{5-2\epsilon'}$, leading to the convergence rate

$$d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m(x)) = O_p(n_0^{-\frac{4-\epsilon-\epsilon'}{5-2\epsilon'}}),$$

which is faster than the convergence rate of local Fréchet regression $O_p(n_0^{-4/5})$ [26].

E Transfer Learning for Wasserstein Regression with Empirical Measures

Fréchet regression assumes that each distribution is fully observed. However, this assumption is often impractical, since in real-world settings one rarely encounters datasets where each observation itself constitutes a full distribution. To address, Wasserstein regression with empirical distributions has been proposed [43], where distributions are replaced by their empirical versions, reaching a convergence rate of $O_p(n^{-1/2} + N_{\min}^{-1/4})$, where $N_{\min}$ representing the minimum number of observations. In this section we do the same modification to handle this obstacle, providing algorithms for global and local Wasserstein transfer learning with empirical measures (WaTL-EM), respectively. The only difference between them and those with full distributions is that we substitute empirical distributions $\hat{\nu}_i^{(k)}$ defined as $\frac{1}{N_i} \sum_{j=1}^{N_i^{(k)}} \delta_{y_{ij}^{(k)}}$ for the true but unobservable distributions $\nu_i^{(k)}$, $1 \leq i \leq n_k$, $1 \leq k \leq K$,

where we recall the settings in Subsection 3.2 and Appendix D, additionally assume $\{y_{ij}^{(k)}\}_{j=1}^{N_i^{(k)}}$ are independently sampled from $\nu_i^{(k)}$ and define δ_z as the Dirac measure at z . In addition, we denote $N_{\min}$ as $\min_{i,k} N_i^{(k)}$.

Theorem 6. Assume Conditions 1 and 2 hold and the regularization parameter satisfies $\lambda \asymp (n_0^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon}$ for some $\epsilon > 0$. Then, for the WaTL-EM algorithm and a fixed $x \in \mathbb{R}^p$, it holds that

$$d_{\mathcal{W}}^2(\hat{m}_{G,EM}^{(0)}(x), m_G^{(0)}(x)) = O_p \left((n_0^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon} \left(\psi + \frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-1/2} + N_{\min}^{-1/4} \right) \right).$$

Algorithm 5 Wasserstein Transfer Learning with Empirical measures (WaTL-EM)

Input: Target and source data $\{(x_i^{(0)}, \{y_{ij}^{(0)}\}_{j=1}^{N_i^{(0)}})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \{y_{ij}^{(k)}\}_{j=1}^{N_i^{(k)}})\}_{i=1}^{n_k})$, regularization parameter λ , and query point $x \in \mathbb{R}^p$.

Output: Target estimator $\hat{m}_{G,EM}^{(0)}(x)$.

- 1: Empirical measures: $\hat{\nu}_i^{(k)} = \frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \delta_{y_{ij}^{(k)}}$.
 - 2: Weighted auxiliary estimator: $\hat{f}_{EM}(x) = \frac{1}{n_0 + n_A} \sum_{k=0}^K n_k \hat{f}_{EM}^{(k)}(x)$, where $\hat{f}_{EM}^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iG}^{(k)}(x) F_{\hat{\nu}_i^{(k)}}^{-1}$.
 - 3: Bias correction using target data: $\hat{f}_{0,EM}(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iG}^{(0)}(x) \|F_{\hat{\nu}_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}_{EM}(x)\|_2$.
 - 4: Projection to Wasserstein space: $\hat{m}_{G,EM}^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_{0,EM}(x)\|_2$.
-

Algorithm 6 Local Wasserstein Transfer Learning with Empirical measures (LWaTL-EM)

Input: Target and source data $\{(x_i^{(0)}, \{y_{ij}^{(0)}\}_{j=1}^{N_i^{(0)}})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \{y_{ij}^{(k)}\}_{j=1}^{N_i^{(k)}})\}_{i=1}^{n_k})$, regularization parameter λ , bandwidth h and query point $x \in \mathbb{R}$.

Output: Target estimator $\hat{m}_{L,h}^{(0)}(x)$.

- 1: Empirical measures: $\hat{\nu}_i^{(k)} = \frac{1}{N_i^{(k)}} \sum_{j=1}^{N_i^{(k)}} \delta_{y_{ij}^{(k)}}$.
- 2: Weighted auxiliary estimator

$$\hat{f}_{h,EM}(x) = \frac{1}{n_0 + n_A} \sum_{k=0}^K n_k \hat{f}_{h,EM}^{(k)}(x),$$

where $\hat{f}_{h,EM}^{(k)}(x) = n_k^{-1} \sum_{i=1}^{n_k} s_{iL}^{(k)}(x, h) F_{\hat{\nu}_i^{(k)}}^{-1}$ and $n_A = \sum_{k=1}^K n_k$.

- 3: Bias correction using target data

$$\hat{f}_{0h,EM}(x) = \arg \min_{g \in L^2(0,1)} \frac{1}{n_0} \sum_{i=1}^{n_0} s_{iL}^{(0)}(x, h) \|F_{\hat{\nu}_i^{(0)}}^{-1} - g\|_2^2 + \lambda \|g - \hat{f}_{h,EM}(x)\|_2.$$

- 4: Projection to Wasserstein space

$$\hat{m}_{L,h,EM}^{(0)}(x) = \arg \min_{\mu \in \mathcal{W}} \|F_\mu^{-1} - \hat{f}_{0h,EM}(x)\|_2.$$

where $\psi = \max_{1 \leq k \leq K} \|f^{(0)}(x) - f^{(k)}(x)\|_2$ quantifies the maximum discrepancy between the target and source.

Proof. First note the following result has been established in [43]. $\|\hat{f}_{EM}^{(k)}(x) - f^{(k)}(x)\|_2 = O_p(n_k^{-1/2} + N_{\min}^{-1/4})$.

Also, an analogy of Theorem 1 can be established in a similar way, just note that analysis in proof of 3 we have established that

$$\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x)| = O_p(1).$$

Then using Theorem 7.9 in [3], we have

$$\begin{aligned} & E \left[\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x)| \|F_{\nu_i^{(k)}}^{-1} - F_{\hat{\nu}_i^{(k)}}^{-1}\|_2 \mid \{(x_i^{(0)}, \{y_{ij}^{(0)}\}_{j=1}^{N_i^{(0)}})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \{y_{ij}^{(k)}\}_{j=1}^{N_i^{(k)}})\}_{i=1}^{n_k}) \right] \\ & \leq \frac{1}{2N_{\min}^{1/4}} \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iG}^{(k)}(x)| = O_p(N_{\min}^{-1/4}). \end{aligned}$$

Hence we have $\|\hat{f} - \hat{f}_{EM}\|_2 = O_p(N_{\min}^{-1/4})$. Then the arguments in the proof of Theorem 2 still hold if $F_{\nu_i^{(0)}}^{-1}$, $\hat{f}^{(0)}$, $\hat{f}$ are replaced by their empirical-measure version. And the rate turns to be $O_p\left((n_0^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon} \left(\psi + \frac{\sum_{k=0}^K \sqrt{n_k}}{n_0 + n_{\mathcal{A}}} + (n_0 + n_{\mathcal{A}})^{-1/2} + N_{\min}^{-1/4}\right)\right)$. $\square$

Theorem 7. Assume Condition 5 holds and the regularization parameter satisfies $\lambda \asymp (n_0^{-1/2} h^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon}$ for some $\epsilon > 0$. Then, for the LWaTL-EM algorithm and a fixed $x \in \mathbb{R}^p$, it holds that

$$\begin{aligned} & d_{\mathcal{W}}^2(\hat{m}_{L,h}^{(0)}(x), m^{(0)}(x)) \\ & = O_p\left((n_0^{-1/2} h^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon} (N_{\min}^{-1/4} \right. \\ & \quad \left. + \psi_L + h^2 + h^{-1/2} \left(\sqrt{\sum_{k=0}^K \frac{1}{n_k}} + \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right) \right), \end{aligned}$$

where $\psi_L = \max_{1 \leq k \leq K} \|f_{\oplus}^{(0)}(x) - f_{\oplus}^{(k)}(x)\|$.

Proof. Using the same arguments in the proof of Theorem 2 in [43] we can show that

$$\|f_h^{(0)}(x) - \hat{f}_{h,EM}^{(0)}(x)\|_2 = O_p(n_0^{-1/2} h^{-1/2} + N_{\min}^{-1/4}).$$

Analysis in the proof of Theorem 4 shows

$$\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h)| = O_p(1).$$

Then using Theorem 7.9 in [3], we have

$$\begin{aligned} & E \left[\frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h)| \|F_{\nu_i^{(k)}}^{-1} - F_{\hat{\nu}_i^{(k)}}^{-1}\|_2 \mid \{(x_i^{(0)}, \{y_{ij}^{(0)}\}_{j=1}^{N_i^{(0)}})\}_{i=1}^{n_0} \cup (\cup_{1 \leq k \leq K} \{(x_i^{(k)}, \{y_{ij}^{(k)}\}_{j=1}^{N_i^{(k)}})\}_{i=1}^{n_k}) \right] \\ & \leq \frac{1}{2N_{\min}^{1/4}} \frac{1}{n} \sum_{k=1}^K \sum_{i=1}^{n_k} |s_{iL}^{(k)}(x, h)| = O_p(N_{\min}^{-1/4}). \end{aligned}$$

Hence we have $\|\hat{f}_h - \hat{f}_{h,EM}\|_2 = O_p(N_{\min}^{-1/4})$. Then the arguments in the proof of Theorem 4 still hold if $F_{\nu_i^{(0)}}^{-1}$, $\hat{f}_h^{(0)}$, $\hat{f}_h$ are replaced by their empirical-measure version. And the rate turns to be $O_p\left((n_0^{-1/2} h^{-1/2} + N_{\min}^{-1/4})^{1-\epsilon} [N_{\min}^{-1/4} + \psi_L + h^2 + h^{-1/2} \left(\sqrt{\sum_{k=0}^K \frac{1}{n_k}} + \sum_{k=0}^K \frac{\sqrt{n_k}}{n_0 + n_{\mathcal{A}}} \right) \right]$. $\square$