
Meta Optimality for Demographic Parity Constrained Regression via Post-Processing

Kazuto Fukuchi^{1 2}

Abstract

We address the regression problem under the constraint of demographic parity, a commonly used fairness definition. Recent studies have revealed fair minimax optimal regression algorithms, the most accurate algorithms that adhere to the fairness constraint. However, these analyses are tightly coupled with specific data generation models. In this paper, we provide meta-theorems that can be applied to various situations to validate the fair minimax optimality of the corresponding regression algorithms. Furthermore, we demonstrate that fair minimax optimal regression can be achieved through post-processing methods, allowing researchers and practitioners to focus on improving conventional regression techniques, which can then be efficiently adapted for fair regression.

1. Introduction

Machine learning systems have become increasingly prevalent in decision-making across various domains, including healthcare, finance, and criminal justice. While these systems promise more efficient and data-driven decisions, they also raise significant concerns regarding fairness and equity. As machine learning models learn from historical data, they can inadvertently perpetuate or even exacerbate existing societal biases, leading to unfair outcomes for certain social groups. Numerous instances of unfair behaviors in real-world machine learning systems have been reported, including biased recidivism risk prediction (Angwin et al., 2016), discriminatory hiring practices (Dastin, 2018), inequitable facial recognition performance (Crockford, 2020; Najibi, 2020), and biased credit scoring (Vigdor, 2019). These reports underscore the urgent need to address the

issue of unfairness in machine learning.

To mitigate unfair bias, researchers have developed diverse methodologies for constructing accurate predictors that ensure certain fairness definitions. These methodologies can be categorized into three main types: pre-processing techniques that modify the training data (Feldman et al., 2015), in-processing methods that incorporate fairness constraints during model training (Chuang & Mroueh, 2020; Zhang et al., 2018; Du et al., 2021; Cotter et al., 2019; Khalili et al., 2023; Jovanović et al., 2023), and post-processing approaches that adjust the model’s outputs (Menon & Williamson, 2018; Chzhen et al., 2019; Jiang et al., 2020; Schreuder & Chzhen, 2021; Chen et al., 2023; Xu & Strohmer, 2023; Xian et al., 2023; Zhao & Gordon, 2019; Chzhen et al., 2020). Methodologies of each type have been proposed for various fairness definitions, including demographic parity (Pedreshi et al., 2008), equalized odds (Hardt et al., 2016), multicalibration (Kleinberg et al., 2017), and individual fairness (Dwork et al., 2012). Each of these fairness criteria aims to address different aspects of algorithmic bias. These diverse methodologies have significantly advanced the field of fair machine learning, enabling researchers and practitioners to address fairness concerns in various contexts and applications.

Recent advancements in fair learning algorithms have led to the development of methods that achieve the best possible predictive accuracy while adhering to specific fairness definitions, particularly demographic parity. Such advancements were achieved by constructing fair learning algorithms and proving their minimax optimality under regression (Chzhen & Schreuder, 2022; Fukuchi & Sakuma, 2023) and classification (Zeng et al., 2024) setups. The fair minimax optimal algorithm is an algorithm that satisfies the fairness definition and minimizes the worst-case error taken over a certain set of data generation models. No fair algorithm can outperform the fair minimax optimal algorithm in the sense of the worst-case error, as the minimization is taken over all the fair algorithms.

While these optimal methods represent significant progress, they are often tightly coupled with specific data generation models, limiting their applicability to a broader range of real-world scenarios. For example, Chzhen & Schreuder

¹Institute of Systems and Information Engineering, University of Tsukuba, Ibaraki, Japan ²RIKEN AIP, Tokyo, Japan. Correspondence to: Kazuto Fukuchi <fukuchi@cs.tsukuba.ac.jp>.

(2022) and Fukuchi & Sakuma (2023) each employ certain linear models of the outcome with Gaussian features, albeit with different specific formulations. Zeng et al. (2024) work under assumptions including that the regression function is within the Hölder class and that both the margin and strong density conditions are satisfied. The dependence on these particular data distributions and model assumptions can restrict the applicability of these approaches, potentially hindering their adoption in diverse applications where the underlying data characteristics may differ significantly from these specific conditions.

Our contributions (Meta-optimality) We address the limitations of existing analyses for fair regression by establishing a meta-theorem that applies to a wide range of scenarios. This meta-theorem provides a connection between the minimax optimal error for fair regression and that for conventional regression, allowing for rates tailored to various situations by leveraging well-established results regarding the minimax optimality in conventional regression. Our approach can combine with minimax optimal regressions under diverse smoothness assumptions (e.g., Hölder, Sobolev, and Besov spaces (Donoho & Johnstone, 1998; Giné & Nickl, 2015)) and minimax optimal deep learning methods (Schmidt-Hieber, 2020; Suzuki, 2018; Suzuki & Nitanda, 2021; Nishimura & Suzuki, 2023).

(Optimal fair regression by post-processing) We propose a post-processing algorithm that leverages an optimal conventional regression algorithm. Guided by our meta-theorem, this construction ensures minimax optimality under the assumptions employed by the conventional regression algorithm. Since the proposed algorithm is post-processing, practitioners can concentrate on refining conventional regression methods, which can then be seamlessly adapted for fair regression.

(Convergence rate analysis for optimal transport map estimation in Wasserstein barycenters) A key component of our algorithm is optimal transport map estimation within the Wasserstein barycenter problem, which seeks a distribution (often called the barycenter) that minimizes the (Wasserstein) distances to a set of distributions. The optimal transport map is a mapping between distributions that achieves the minimum cost. One of our main contributions is to provide a convergence rate analysis of a transport map estimator for the Wasserstein barycenter problem, which may be of independent interest. The analyzed estimator is based on Korotin et al. (2020), but they did not provide a convergence rate analysis. The detailed discussion will appear in Section 7.

All the missing proofs can be found in Appendix A.

Notations For a positive integer m , let $[m] = \{1, \dots, m\}$ and Δ_m denote the probability simplex over $[m]$. We denote the indicator function by $\mathbb{1}$. For real values a and b , we define $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For a sequence a_t indexed by $t \in \mathcal{T}$, we represent the family $(a_t)_{t \in \mathcal{T}}$ as a_\cdot . The first derivative of a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is denoted by Df . Given an event $\mathcal{E} \in \mathfrak{Z}$ in a probability space $(\mathcal{Z}, \mathfrak{Z}, \nu)$, we denote its complement by \mathcal{E}^c and its probability by $\mathbb{P}_\nu\{\mathcal{E}\}$. For a random variable X from $(\mathcal{Z}, \mathfrak{Z}, \nu)$ to a measurable space $(\mathcal{X}, \mathfrak{X})$, we denote its expectation and variance by $\mathbb{E}_\nu[X]$ and $\mathbb{V}_\nu[X]$, respectively.

2. Problem Setup and Preliminaries

2.1. Fair Regression Problems

Consider a fair regression problem with $M \geq 2$ social groups. Let \mathcal{X} and $\Omega \subset \mathbb{R}$ be the domains of features and outcomes, respectively, where we assume Ω is open and bounded. For each social group $s \in [M]$ (e.g., male and female for gender), let $X^{(s)} \in \mathcal{X}$ and $Y^{(s)} \in \Omega$ be random variables on a probability measure space $(\mathcal{Z}, \mathfrak{Z}, \mu_s)$, representing the features and outcomes of an individual in group s , respectively. The goal of the regression problem is to construct a (group-wise) regressor f_\cdot , mappings from \mathcal{X} to Ω indexed by $s \in [M]$, that accurately predicts $Y^{(s)}$ based on $X^{(s)}$. The ideal regressor, known as the Bayes-optimal regressor, is defined as $f_{\mu_s}^* = \arg \min_f \mathbb{E}_{\mu_s}[(f(X^{(s)}) - Y^{(s)})^2]$ and is given by $f_{\mu_s}^*(X^{(s)}) = \mathbb{E}_{\mu_s}[Y^{(s)} | X^{(s)}]$. We use $\mu_{Y,s}$ and $\mu_{X,s}$ to denote the laws of $Y^{(s)}$ and $X^{(s)}$, respectively. Additionally, we denote the law of $f_{\mu_s}^*(X^{(s)})$ by $\mu_{f,s}$.

Given samples consisting of n_s i.i.d. copies of $(X^{(s)}, Y^{(s)})$, the objective of the learning algorithm is to construct a regressor $\tilde{f}_{n,\cdot}$ that maximizes accuracy while satisfying a fairness constraint. Let $n = \sum_{s \in [M]} n_s$ for notational convenience. We now introduce the definition of fairness, define a measure of accuracy, and provide the definition of the fair optimal algorithm.

Fairness We employ demographic parity (Pedreshi et al., 2008) as our fairness criterion. A regressor f_\cdot satisfies demographic parity if its output distribution remains invariant across all groups $s \in [M]$.

Definition 2.1. A regressor f_\cdot satisfies (strict) demographic parity if, for all $s, s' \in [M]$ and for all events E , $\mathbb{P}_{\mu_s}\{f_s(X^{(s)}) \in E\} = \mathbb{P}_{\mu_{s'}}\{f_{s'}(X^{(s')}) \in E\}$.

Let $\bar{\mathcal{F}}(\mu_{X,\cdot})$ denote the set of all regressors satisfying demographic parity for given laws $\mu_{X,\cdot}$.

Instead of enforcing strict demographic parity in Definition 2.1, we adopt the concept of *fairness consistency* (Chzheng et al., 2020; Fukuchi & Sakuma, 2023).

Definition 2.2. A learning algorithm is consistently fair if $\bar{f}_{n,:}$ converges in probability to an element of $\bar{\mathcal{F}}(\mu_{X,:})$ as n_1, \dots, n_M approach infinity.

Definition 2.2 implies that a consistently fair learning algorithm eventually constructs a regressor that satisfies strict demographic parity given a sufficiently large sample size.

Accuracy We evaluate the accuracy of a given regressor by measuring its expected squared distance from the fair Bayes-optimal regressor. Given probability measures ν_i on a measurable space $(\mathcal{Z}, \mathfrak{Z})$ indexed by $[M]$ and weights $w_i \in \Delta_M$ (known to the learner), we define the squared distance d_{ν_i} between functions $f_s, f'_s : \mathcal{Z} \rightarrow \mathbb{R}$ indexed by $s \in [M]$ as

$$\begin{aligned} d_{\nu_i}^2(f_i, f'_i) &= \sum_{s \in [M]} w_s \int (f_s(z) - f'_s(z))^2 \mu_s(dz) \\ &:= \sum_{s \in [M]} w_s d_{\nu_s}^2(f_s, f'_s). \end{aligned}$$

The fair Bayes-optimal regressor is defined as the regressor that satisfies strict demographic parity and minimizes the deviation from the Bayes-optimal regressor:

$$\bar{f}_{\mu_i}^* = \arg \min_{f_i \in \bar{\mathcal{F}}(\mu_{X,i})} d_{\mu_{X,i}}^2(f_i, f_{\mu_i}^*).$$

The accuracy of a regressor f_i is then evaluated by $d_{\mu_{X,i}}^2(f_i, \bar{f}_{\mu_i}^*)$.

Optimality The fair minimax optimal algorithm for a set of distributions \mathcal{P} is a consistently fair algorithm that achieves the fair minimax optimal error over \mathcal{P} . The fair minimax optimal error over \mathcal{P} is defined as

$$\bar{\mathcal{E}}_n(\mathcal{P}) = \inf_{\bar{f}_{n,:} : \text{fair}} \sup_{\mu_i \in \mathcal{P}} \mathbb{E}_{\mu_i^n} [d_{\mu_{X,i}}^2(\bar{f}_{n,:}, \bar{f}_{\mu_i}^*)], \quad (1)$$

where the infimum is taken over all consistently fair learning algorithms, and $\mathbb{E}_{\mu_i^n}$ denotes the expectation over samples. Thus, no consistently fair learning algorithm can outperform the fair minimax optimal algorithm in terms of worst-case expected deviation.

2.2. Fair Bayes-Optimal Regressors and Optimal Transport Maps

Recent analyses have characterized fair Bayes-optimal regressors using optimal transport maps that arise in the Wasserstein barycenter problem (Chzhen et al., 2020; Chzhen & Schreuder, 2022). Given two probability measures ν and ν' , the optimal transport map with a quadratic cost function is the unique solution of Monge's formulation of the optimal transportation problem between ν and ν' , i.e.,

a transport map $\vartheta^* : \mathbb{R} \rightarrow \mathbb{R}$ that realizes the infimum

$$W_2^2(\nu, \nu') = \inf_{\vartheta : \vartheta_{\#}\nu = \nu'} \int \frac{1}{2} (z - \vartheta(z))^2 \nu(dz),$$

where $\vartheta_{\#}\nu$ denotes the pushforward measure of ν by ϑ ¹. Given probability measures ν_1, \dots, ν_k , the Wasserstein barycenter problem with weights $w_i \in \Delta_k$ is defined as

$$\inf_{\nu} \sum_{i \in [k]} w_i W_2^2(\nu_i, \nu). \quad (2)$$

We refer to the unique solution of Equation (2) as the barycenter of ν_i with weights w_i . The optimal transport map from ν_i to the barycenter of ν_i is denoted by $\vartheta_{\nu_i}^*$. Building on these concepts, the fair Bayes-optimal regressor is obtained as follows:

Theorem 2.3 (Chzhen et al. (2020)). Assume that $\mu_{f,s}$ admits a density for all $s \in [M]$. Then, the fair Bayes-optimal regressor is given by

$$\bar{f}_{\mu_i,s}^*(x) = (\vartheta_{\mu_{f,s}}^* \circ f_{\mu_i,s}^*)(x).$$

Theorem 2.3 reveals that the fair Bayes-optimal regressor is characterized by the Bayes-optimal regressor $f_{\mu_i}^*$ and the optimal transport maps $\vartheta_{\mu_{f,i}}^*$. Throughout the paper, we assume $\mu_{f,s}$ admits a density for all $\mu_i \in \mathcal{P}$ and $s \in [M]$ so that the condition of Theorem 2.3 holds.

2.3. Potential Minimization and Optimal Transport Maps

The transport maps in Wasserstein barycenter problems are characterized by the minimizer of multiple correlation over congruent potentials (Korotin et al., 2020). Congruent potentials with weights $w_i \in \Delta_k$ are convex and lower semi-continuous functions $u : \mathbb{R} \rightarrow \mathbb{R}$ such that $\sum_{i \in [k]} w_i Du^\dagger(z) = z$ for all $z \in \Omega$, where $u^\dagger(z) = \sup_x (zx - u(x))$ is the convex conjugate of u . Given probability measures ν_i , the multiple correlation of ν_i with weights w_i for congruent potentials u_i is defined as

$$C(u_i; \nu_i) = \sum_{i \in [k]} w_i \int u_i d\nu_i.$$

Let u_i^* denote the optimal congruent potentials that minimize $C(u_i; \nu_i)$. Through the analyses by Agueh & Carlier (2011); Álvarez Esteban et al. (2016), we obtain the following characterization of the optimal transport maps:

Corollary 2.4 (Agueh & Carlier (2011); Álvarez Esteban et al. (2016)). Let ϑ_i^* be the optimal transport maps from ν_i to the barycenter of ν_i with weights w_i . Suppose that for all $i \in [k]$, ν_i admits densities. Then, $\vartheta_i^* = Du_i^*$ for $i \in [k]$.

¹We use the notation W_2 because W_2 is known as the 2-Wasserstein distance.

By [Corollary 2.4](#), we can obtain the optimal transport maps ϑ^* by solving $\inf_{u, \nu: \text{congruent}} C(u, \nu)$.

3. Main Result

Our main result is a meta-theorem that characterizes the fair minimax optimal error (see [Equation \(1\)](#)), showing how its convergence rate depends on \mathcal{P} . We begin by introducing several technical assumptions on \mathcal{P} before presenting our main meta-theorem.

For convenience, we introduce several notations. We use notations $\mathcal{P}_Y = \{\mu_{Y, \cdot} : \mu \in \mathcal{P}\}$, $\mathcal{P}_X = \{\mu_{X, \cdot} : \mu \in \mathcal{P}\}$, and $\mathcal{P}_f = \{\mu_{f, \cdot} : \mu \in \mathcal{P}\}$. We denote $\mathcal{P}_s = \{\mu_s : \mu \in \mathcal{P}\}$ for $s \in [M]$. Let $\mathcal{F}_{\mathcal{P}} = \{f_{\mu, \cdot}^* : \mu \in \mathcal{P}\}$, and let $\mathcal{F}_{\mathcal{P}, s} = \{f_{\mu, s}^* : \mu \in \mathcal{P}\}$ for $s \in [M]$. Let $\Theta_{\mathcal{P}} = \{\vartheta_{\mu, \cdot}^* : \mu \in \mathcal{P}\}$. We omit \mathcal{P} in the subscript of these notations if \mathcal{P} is clear from the context.

Assumptions We begin with our first assumption.

Assumption 3.1. \mathcal{P} satisfies the following three conditions:

1. $\mathcal{P}_Y \times \mathcal{P}_X \subseteq \{(\mu_{Y, \cdot}, \mu_{X, \cdot}) : \mu \in \mathcal{P}\}$,
2. for any permutation π over $[M]$, $\mu_{\pi(\cdot)} \in \mathcal{P}$ if $\mu \in \mathcal{P}$,
3. \mathcal{F}_s is convex, meaning for any $f, f' \in \mathcal{F}_s$ and $t \in (0, 1)$, $tf + (1-t)f' \in \mathcal{F}_s$.

Intuitively, the first and second conditions imply that the learner has no prior knowledge about (i) how the distributions of $Y^{(s)}$ and $X^{(s)}$ are related, nor (ii) how the distributions $\{\mu_s\}_{s \in [M]}$ differ. These conditions are naturally satisfied in many real-world scenarios. Note that the third condition does not require the regression functions themselves to be convex but rather that the set \mathcal{F}_s is convex, which can still accommodate non-convex functions.

Next, we impose assumptions on $\mu_{f, \cdot}$ to facilitate the estimation of the optimal transport maps $\vartheta_{\mu_{f, \cdot}}^*$. First, we assume that $\vartheta_{\mu_{f, \cdot}}^*$ are elements of Lipschitz and strictly increasing functions \mathcal{M}_L ². Specifically, \mathcal{M}_L is defined as the set of functions $\vartheta : \Omega \rightarrow \Omega$ satisfying

$$L^{-1}(y - x) \leq \vartheta(x) - \vartheta(y) \leq L(y - x) \quad \forall x > y. \quad (3)$$

Second, we assume that $\mu_{f, s}$ satisfies the Poincaré-type inequality: there exists a constant $C_P > 0$ such that for any function $g : \Omega \rightarrow \Omega$ with an L -Lipschitz continuous gradient,

$$\mathbb{V}_{\mu_s} [g(X^{(s)})] \leq C_P \mathbb{E}_{\mu_s} \left[\left(Dg(X^{(s)}) \right)^2 \right]. \quad (4)$$

Assumption 3.2. There exists a constant $L > 1$ such that for all $\mu \in \mathcal{P}$ and all $s \in [M]$, 1) $\vartheta_{\mu_{f, s}}^* \in \mathcal{M}_L$ and 2) $\mu_{f, s}$ satisfies the Poincaré-type inequality in [Equation \(4\)](#).

²It is worth noting that any transport map over the real line is a non-decreasing function due to the convexity of the potential.

Note that similar assumptions to [Assumption 3.2](#) are also employed in studies on transport map estimation, including [Hütter & Rigollet \(2021\)](#); [Divol et al. \(2024\)](#).

We use the following complexity measure of the class of transport maps Θ based on metric entropy. Given $\epsilon > 0$, the ϵ -covering number of a set $A \subseteq \mathcal{X}$ with a metric space (\mathcal{X}, d) is denoted as $N(\epsilon, A, d)$; namely, $N(\epsilon, A, d)$ denotes the minimum number of balls whose union covers A . Let $\ln_+(x) = 0 \vee \ln(x)$. The complexity measure is defined as follows:

Definition 3.3. The complexity of \mathcal{P}_f is (α, β) for $\alpha > 0$ and $\beta \geq 0$ if there exist constants $C, C' > 0$ and $\bar{\epsilon} > 0$, and a sequence of subsets $\Theta_1 \subseteq \Theta_2 \subseteq \dots \subseteq \{\vartheta_{\nu, \cdot}^* : \nu \in \mathcal{P}_f\}$ such that for any integer $j \geq 0$,

1. $\sup_{\nu \in \mathcal{P}_f} \inf_{\vartheta \in \Theta_j} d_{\nu, \cdot}^2(\vartheta, \vartheta_{\nu, \cdot}^*) \leq C 2^{-\alpha j}$,
2. $\sup_{\nu \in \mathcal{P}_f} \ln N(\epsilon, \Theta_j, d_{\nu, \cdot}) \leq C' 2^{\beta j} \ln_+(1/\epsilon)$ for $\epsilon \in (0, \bar{\epsilon}]$.

In our main theorem, we will assume the complexity is (α, β) for some α and β .

Meta-theorem We now characterize the fair minimax optimal error $\mathcal{E}_n(\mathcal{P})$ in terms of the conventional minimax optimal error. Specifically, for a given group $s \in [M]$, the group s 's conventional minimax optimal error is defined as

$$\mathcal{E}_k(\mathcal{P}_s) = \inf_{f_k} \sup_{\mu_s \in \mathcal{P}_s} \mathbb{E}_{\mu_s^k} \left[d_{\mu_{X, s}}^2(f_k, f_{\mu, s}^*) \right]$$

where the infimum is taken over all regression algorithms that take k i.i.d. copies of $(X^{(s)}, Y^{(s)})$ as the observed sample. Let $\tilde{n} = \min_{s \in [M]} n_s / w_s$.

Theorem 3.4. Assume [Assumptions 3.1](#) and [3.2](#) and that the complexity of \mathcal{P}_f is (α, β) for some $\alpha > 0$ and $\beta \geq 0$. Then, there exists a consistently fair learning algorithm such that

$$\mathcal{E}_n(\mathcal{P}_1) \leq \bar{\mathcal{E}}_n(\mathcal{P}) \leq C \cdot$$

$$\left(L^2 \sum_{s \in [M]} w_s \mathcal{E}_{n_s}(\mathcal{P}_s) + \left(\frac{\tilde{n}}{\ln(\tilde{n})} \right)^{-\alpha/(\alpha+\beta)} \right),$$

for some constant $C > 0$.

We highlight several implications of [Theorem 3.4](#):

1. Assuming there exists a constant $c > 0$ such that $n_s \geq cw_s n$ for all $s \in [M]$, [Theorem 3.4](#) shows that the optimal rate with respect to n is $\mathcal{E}_n(\mathcal{P}_1)$ (note that $\mathcal{E}_n(\mathcal{P}_1) = \dots = \mathcal{E}_n(\mathcal{P}_M)$) whenever $\mathcal{E}_n(\mathcal{P}_1)$ is larger than $(\frac{n}{\ln(n)})^{-\alpha/(\alpha+\beta)}$. In such cases, the rate of $\bar{\mathcal{E}}_n(\mathcal{P})$ can vary with \mathcal{P} along with $\mathcal{E}_n(\mathcal{P}_1)$.
2. In general, [Theorem 3.4](#) implies that if the conventional regression problem is more difficult than the transport map estimation problem, then the optimal fair regression

error is dominated by the conventional minimax error. This situation commonly arises in high-dimensional settings for \mathcal{X} (e.g., image, text, or audio regression).

3. [Chzhen et al. \(2020\)](#) also established a similar upper bound on the error under the demographic parity constraint. However, their result has two notable limitations. First, their analysis requires stronger assumptions than ours. For instance, they assume that the conventional regression algorithm admits a sub-Gaussian high-probability error bound, whereas our results only require a bound on the expected squared error. Furthermore, while they require uniform upper and lower bounds on the density of $\mu_{f,s}$, we only assume the Poincaré-type inequality. These differences broaden the applicability of our meta-theorem relative to their findings.
4. The second limitation of [Chzhen et al. \(2020\)](#) is that their results cannot achieve a convergence rate faster than $n^{-1/2}$, since their upper bound on the estimation error of $\vartheta_{\mu,\cdot}^*$ is $n^{-1/2}$ and dominates the other terms. In contrast, our result can achieve a rate faster than $n^{-1/2}$ by exploiting the smoothness structure of $\vartheta_{\mu,\cdot}^*$.

Illustrative Example To concretely demonstrate the implications of our theoretical results, we consider a representative scenario in which the regression function $f_{\mu,s}^*$ is a composition of multiple functions, as studied by [Schmidt-Hieber \(2020\)](#), and the optimal transport map $\vartheta_{\mu,s}^*$ lies within a Sobolev function class. Specifically, let $f_{\mu,s}^*$ belong to the class

$$\left\{ g_q \circ \dots \circ g_0 : \right. \\ \left. g_i = (g_{ij})_j : [a_i, b_i]^{d_i} \rightarrow [a_{i+1}, b_{i+1}]^{d_{i+1}}, \right. \\ \left. g_{ij} \in C_{t_i}^{\beta_i}([a_i, b_i]^{t_i}) \right\},$$

where C_r^β denotes the Hölder class of functions with smoothness parameter β and r -dimensional input, d_i is the input dimension of g_i , and $t_i < d_i$ indicates that each g_{ij} depends on only t_i out of d_i variables. This structure captures the notion of compositional functions with sparse dependencies, which is prevalent in high-dimensional statistical learning.

For regression functions of this form, the minimax optimal error for group s satisfies $\mathcal{E}_n(\mathcal{P}_s) = \Theta\left(\max_i n^{-\beta_i^*/(2\beta_i^*+t_i)}\right)$ up to logarithmic factors, where $\beta_i^* = \beta_i \prod_{\ell=i+1}^q (\beta_\ell \wedge 1)$. Also, for the class of transport maps $\vartheta_{\mu,s}^*$ taken to be the Sobolev class of smoothness $\gamma > 0$, [Definition 3.3](#) is satisfied with $\alpha = 2\gamma$ and $\beta = 1$ by choosing Θ_j as the span of the first j wavelet basis functions. Consequently, the fair minimax error is obtained as

$$\bar{\mathcal{E}}_n(\mathcal{P}) = \Theta\left(\max_i n^{-\beta_i^*/(2\beta_i^*+t_i)} + n^{-2\gamma/(2\gamma+1)}\right),$$

again up to logarithmic factors, provided that there exists a constant $c > 0$ such that $n_s \geq c w_s n$ for all $s \in [M]$.

An important insight from this example is that if the smoothness of the regression components and the transport maps are comparable (i.e., $\beta_i^* \approx \gamma$), then the minimax error is dominated by the regression term whenever $t_i > 1$ for some i . Here, t_i reflects the intrinsic dimensionality of the essential intermediate data representation. In practical situations, the intermediate representations may be multi-dimensional ($t_i > 1$), and thus the overall rate is determined by the conventional regression problem.

4. Optimal Algorithm and Upper Bound

In this section, we present our fair regression algorithm. Its core structure is similar to algorithms proposed by [Chzhen et al. \(2020\)](#); [Chzhen & Schreuder \(2022\)](#). Leveraging [Theorem 2.3](#), we obtain the fair Bayes-optimal regressor by composing the optimal transport maps $\vartheta_{\mu_{f,\cdot}}^*$ from $\mu_{f,\cdot}$ to the barycenter of $\mu_{f,\cdot}$ with the conventional Bayes-optimal regressor $f_{\mu,\cdot}^*$. Let $\hat{f}_{n,\cdot}$ and $\vartheta_{n,\cdot}$ be estimators of $f_{\mu,\cdot}^*$ and $\vartheta_{\mu_{f,\cdot}}^*$, respectively. We define the estimator of $\hat{f}_{\mu,\cdot}^*$ as $\hat{f}_{n,s}(x) = (\vartheta_{n,s} \circ \hat{f}_{n,s})(x)$. This procedure can be viewed as post-processing since we first construct $\hat{f}_{n,\cdot}$ using a conventional learning algorithm and then refine its outputs using $\vartheta_{n,\cdot}$.

We employ a minimax-optimal conventional regression algorithm for the model \mathcal{P} as the estimator $\hat{f}_{n,\cdot}$. Our main methodological contribution, distinguishing our work from earlier approaches, is the introduction of a suitable estimator for $\vartheta_{n,\cdot}$. To estimate $\vartheta_{\mu_{f,\cdot}}^*$, we utilize the strategy proposed by [Korotin et al. \(2020\)](#) (see [Section 2.3](#)). We also provide a novel analysis of the convergence rate for the estimation error of $\vartheta_{n,\cdot}$, which was not addressed in [Korotin et al. \(2020\)](#).

We first describe the construction of $\vartheta_{n,\cdot}$ and then demonstrate the overall procedure of our fair regression algorithm along with analyses of its accuracy and fairness.

Barycenter estimation As discussed in [Section 2.3](#), one can obtain the optimal transport maps in the Wasserstein barycenter problem by finding congruent potentials that minimize the multiple correlation $C(u, \nu)$. In the fair regression setting, our goal is to estimate the transport maps in the Wasserstein barycenter problem for $\mu_{f,\cdot}$. However, the learner cannot directly observe $\mu_{f,\cdot}$ since neither $f_{\mu,\cdot}^*$ nor $\mu_{X,\cdot}$ is known. Instead, we substitute $\hat{f}_{n,\cdot}$ for $f_{\mu,\cdot}^*$ and use empirical measures from the observed samples in place of $\mu_{X,\cdot}$. Specifically, let $\mu_{\hat{f},s}$ be the law of $\hat{f}_{n,s}(X^{(s)})$ for $s \in [M]$, and let $\mu_{n,\hat{f},s}$ be the corresponding empirical measure induced by a sample of size n_s . Then, the estimator

Algorithm 1 Optimal fair regression

input Samples $(Y_1^{(s)}, X_1^{(s)}), \dots, (Y_{2n}^{(s)}, X_{2n}^{(s)})$
output $\bar{f}_{n,:}$

- 1: Construct $f_{n,:}$ using a minimax optimal conventional regression algorithm with halves of samples $(Y_1^{(s)}, X_1^{(s)}), \dots, (Y_{n_s}^{(s)}, X_{n_s}^{(s)})$
- 2: Construct $\vartheta_{n,:}$ by Equation (5) with the remaining samples $f_{n,s}(X_{n+1}^{(s)}), \dots, f_{n,s}(X_{2n}^{(s)})$
- 3: $\bar{f}_{n,s}(x) = (\vartheta_{n,s} \circ f_{n,s})(x)$

$\vartheta_{n,:}$ is defined as the minimizer of

$$\inf_{\vartheta_{n,:} \in \Theta_j} C(u_{\vartheta_{n,:}}, \mu_{n,\hat{f},:}), \quad (5)$$

where $u_{\vartheta_{n,:}}$ denotes the potential functions corresponding to the transport maps $\vartheta_{n,:}$, and Θ_j is a sequence of subsets of Θ described in Definition 3.3.

Remark 4.1. We require a specific construction of $u_{\vartheta_{n,:}}$ for technical reasons. We extend the input and output domains of any function $\vartheta : \Omega \rightarrow \Omega$ to \mathbb{R} while preserving the property in Equation (3) by defining $\vartheta(z) = z + c_{\text{sup}}$ for large z and $\vartheta(z) = z + c_{\text{inf}}$ for small z with appropriate constants $c_{\text{sup}}, c_{\text{inf}} \in \mathbb{R}$. We interpret functions in Θ as their extended versions. Letting $u_{\vartheta,s}^\dagger(z) = \int_0^z \vartheta_s^{-1}(x)dx$, we define $u_{\vartheta,s}$ as the convex conjugate of $u_{\vartheta,s}^\dagger$; i.e., $u_{\vartheta,s}(z) = \sup_{x \in \mathbb{R}} (xz - u_{\vartheta,s}^\dagger(x))$.

Overall algorithm Algorithm 1 summarizes the overall procedure of our fair regression algorithm. For simplicity, we assume that the sample size for group s is $2n_s$. In the first step, we execute a minimax-optimal conventional regression algorithm with half of the samples to obtain $f_{n,:}$. By definition, $f_{n,s}$ achieves an error of $\mathcal{E}_n(\mathcal{P}_s)$. In the second step, we estimate the transport map $\vartheta_{n,:}$ via Equation (5) using the remaining samples. As shown in the next corollary, Algorithm 1 achieves the desired properties.

Corollary 4.2. *Under the same conditions as in Theorem 3.4, Algorithm 1 achieves the upper bound in Theorem 3.4 and is consistently fair.*

Remark 4.3 (Limitation). The minimization over congruent potentials may present computational challenges. Korotin et al. (2020) proposed adding a penalty term to enforce congruency instead of handling the constraint directly, which may be more practical. Analysis of convergence rates under approximate satisfaction of congruency remains an important direction for future work.

Connection between fair regression and transport maps estimation in Wasserstein barycenter To prove Corollary 4.2, we relate the regression error and unfairness of

Algorithm 1 to the estimation error of the estimated transport maps $\vartheta_{n,:}$. Specifically, we demonstrate the connection of Algorithm 1's error and unfairness with $d_{\mu_{\hat{f},:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*)$ through the following propositions:

Proposition 4.4. *Let $\bar{f}_{n,:}$ be a regressor obtained by Algorithm 1. Under the same conditions as in Theorem 3.4, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}_{\mu_{X,:}^{2n}} \left[d_{\mu_{X,:}}^2(\bar{f}_{n,:}, \bar{f}_{\mu_{X,:}}^*) \right] \leq C \left(L^2 \sum_{s \in [M]} w_s \mathcal{E}_{n_s}(\mathcal{P}_s) + \mathbb{E}_{\mu_{\hat{f},:}^n} \left[d_{\mu_{\hat{f},:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) \right] \right),$$

where $\mathbb{E}_{\mu_{\hat{f},:}^n}$ denotes the expectation over the samples used for constructing $\vartheta_{n,:}$.

Proposition 4.5. *Let $\bar{f}_{n,:}$ be a regressor obtained by Algorithm 1. Under the same conditions as in Theorem 3.4, we have*

$$\inf_{\nu} \max_{s \in [M]} W_2(\bar{f}_{n,s}, \mu_{X,s}, \nu) \leq \sqrt{\frac{1}{M w_{\min}}} d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) \text{ a.s.,}$$

where $w_{\min} = \min_{s \in [M]} w_s$.

By Proposition 4.4, we obtain an upper bound on the regression error by deriving an upper bound on $\mathbb{E}_{\mu_{\hat{f},:}^n} [d_{\mu_{\hat{f},:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*)]$. Additionally, Proposition 4.5 implies that Algorithm 1 achieves fairness consistency in Definition 2.2 if $\mathbb{P}_{\mu_{\hat{f},:}^n} \left\{ d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) = o(1) \right\} = 1 - o(1)$. We will provide upper bounds on the error $d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*)$ through the analyses shown in the next section.

5. Transport Maps Estimation in Wasserstein Barycenter

In this section, we investigate the convergence rate of our estimator for transport maps estimation in the Wasserstein barycenter. We conduct our analyses under the general setup of transport maps estimation in the Wasserstein barycenter for real-valued probability measures. We first describe the general setup and demonstrate the convergence rate of our estimator, which is the main result of this section. We then provide detailed analyses to support the convergence rate.

Setup Consider the problem of estimating the transport maps that arise in the Wasserstein barycenter problem. Let $\nu_s \in \mathcal{Q}$ be probability measures indexed by $[M]$ on a measurable space (Ω, \mathcal{F}) . Recall that the Wasserstein barycenter problem involves finding the minimizer of the following optimization problem:

$$\inf_{\nu} \sum_{s \in [M]} w_s W_2^2(\nu_s, \nu).$$

Let ν be the minimizer of the above optimization problem, i.e., the Wasserstein barycenter. Recall that we denote $\vartheta_{\nu,s}^*$ as the optimal transport map from ν_s to ν for each $s \in [M]$. Given M samples each i.i.d. from ν_s , the analyst's goal is to estimate $\vartheta_{\nu,\cdot}^*$. We denote the estimator of $\vartheta_{\nu,\cdot}^*$ as $\vartheta_{n,\cdot}$. According to Propositions 4.4 and 4.5, we assess the error of the estimated transport maps by

$$\begin{aligned} d_{\nu}(\vartheta_{n,\cdot}, \vartheta_{\nu,\cdot}^*) &= \sum_{s \in [M]} w_s d_{\nu_s}(\vartheta_{n,s}, \vartheta_{\nu,s}^*) \\ &= \sum_{s \in [M]} w_s \int (\vartheta_{n,s}(z) - \vartheta_{\nu,s}^*(z))^2 \nu_s(dz). \end{aligned}$$

Note that $d_{\nu}(\vartheta_{n,\cdot}, \vartheta_{\nu,\cdot}^*)$ is a random variable with randomness stemming from the samples. The goal of the analyses is thus to provide an upper bound on the expectation of the error or a probabilistic upper bound on the error.

Estimator Our estimator $\vartheta_{n,\cdot}$ is constructed by minimizing the empirical multiple correlation over a sieved subset of transport maps. Let $\Theta = \{\vartheta_{\nu,\cdot}^* : \nu \in \mathcal{Q}\}$ represent the set of all possible transport maps. Denote by $\nu_{n,\cdot}$ the empirical measures corresponding to ν_{\cdot} , as determined by the observed samples. Assume that the complexity of \mathcal{Q} is (α, β) for some $\alpha > 0$ and $\beta \geq 0$, and let $\{\Theta_j\}_j$ be a sequence of subsets of Θ as specified in Definition 3.3. The estimator $\vartheta_{n,\cdot}$ is then defined as the solution to the following optimization problem:

$$\inf_{\vartheta_{\cdot} \in \Theta_j} C(u_{\vartheta_{\cdot},\cdot}, \nu_{n,\cdot}),$$

where j is chosen appropriately.

Estimation Error Bound Our main results in this section are expected and probabilistic upper bounds on $d_{\nu}(\vartheta_{n,\cdot}, \vartheta_{\nu,\cdot}^*)$.

Theorem 5.1. *Let $\vartheta_{n,\cdot} = \arg \min_{\vartheta_{\cdot} \in \Theta_j} C(u_{\vartheta_{\cdot},\cdot}, \nu_{n,\cdot})$ be the estimated transport maps, with j satisfying $2^j \leq (\frac{\tilde{n}}{\ln(\tilde{n})})^{1/(\alpha+\beta)} \leq 2^{j+1}$. Suppose that $\vartheta_{\nu,s}^* \in \mathcal{M}_L$ for some $L > 1$, and that ν_s satisfies the Poincaré-type inequality in Equation (4) for all $\nu \in \mathcal{Q}$. Also, suppose that the complexity of \mathcal{Q} is (α, β) for some $\alpha > 0$ and $\beta \geq 0$. Then, there exists a constant $C > 0$, possibly depending on L and M , such that*

$$\mathbb{E}_{\nu^n} [d_{\nu}^2(\vartheta_{n,\cdot}, \vartheta_{\nu,\cdot}^*)] \leq C \left(\frac{\tilde{n}}{\ln(\tilde{n})} \right)^{-\alpha/(\alpha+\beta)}.$$

Moreover, for all $t \geq 1$, with probability at least $1 - 2e^{-t}$,

$$d_{\nu}^2(\vartheta_{n,\cdot}, \vartheta_{\nu,\cdot}^*) \leq Ct \left(\frac{\tilde{n}}{\ln(\tilde{n})} \right)^{-\alpha/(\alpha+\beta)}.$$

Theorem 5.1 shows that a larger α and a smaller β result in a faster convergence rate.

To establish the upper bound in Theorem 5.1, we follow the convergence rate analysis for the sieved M -estimator (see, e.g., Van Der Vaart & Wellner, 2023), since our estimator in Equation (5) can be regarded as a sieved M -estimator. The main analytical challenge is that the process $\vartheta_{\cdot} \rightarrow C(u_{\vartheta_{\cdot},\cdot}, \nu_{n,\cdot})$ is not a standard empirical process, and thus standard concentration inequalities do not directly apply. In the following subsection, we present the analysis of $\vartheta_{n,\cdot}$, including a concentration inequality for the process $\vartheta_{\cdot} \rightarrow C(u_{\vartheta_{\cdot},\cdot}, \nu_{n,\cdot})$.

5.1. Error Analysis for $\vartheta_{n,\cdot}$

In this subsection, we present the error analysis for $\vartheta_{n,\cdot}$. We begin by formally defining the processes associated with the populational and empirical multiple correlations. The expectation operator E_{ν} is defined as $E_{\nu}u = \mathbb{E}_{\nu}[u(Z)]$ for a function $u : \mathcal{Z} \rightarrow \mathbb{R}$ and $Z \sim \nu$. For a positive integer n , the empirical expectation operator $E_{\nu,n}$ is given by $E_{\nu,n}u = \frac{1}{n} \sum_{i=1}^n u(Z_i)$, where Z_1, \dots, Z_n are drawn i.i.d. from ν . For measurable functions $u_s : \mathcal{Z} \rightarrow \mathbb{R}$ indexed by $[M]$, we define the following operators:

$$E_{\nu}u_{\cdot} = \sum_{s \in [M]} w_s E_{\nu_s} u_s, \quad E_{n,\nu}u_{\cdot} = \sum_{s \in [M]} w_s E_{\nu_s,n} u_s.$$

The populational and empirical multiple correlations are then expressed as

$$C(u_{\vartheta_{\cdot},\cdot}, \nu_{\cdot}) = E_{\nu}u_{\vartheta_{\cdot},\cdot}, \quad C(u_{\vartheta_{\cdot},\cdot}, \nu_{n,\cdot}) = E_{n,\nu}u_{\vartheta_{\cdot},\cdot}.$$

Concentration of the process We now establish the concentration inequality for the process $E_{n,\nu} - E_{\nu}$. We first present a Bernstein-type concentration inequality for fixed potentials $u_{\vartheta_{\cdot},\cdot}$. For $s \in [M]$, let ν_s be a probability measure on a measurable space $(\mathcal{Z}, \mathfrak{F})$. Let $Z^{(s)}$ denote a random variable distributed according to ν_s . Let \mathcal{U} be a class of functions $u_s : \mathcal{Z} \rightarrow \mathbb{R}$ indexed by $s \in [M]$.

Proposition 5.2. *Given $u_{\cdot} \in \mathcal{U}$ such that $\sum_{s \in [M]} w_s \mathbb{V}_{\nu_s}[u_s(Z^{(s)})] \leq \sigma^2$ and $\max_{s \in [M]} |f_s(Z^{(s)}) - E_{\nu_s}u_s| \leq b$ almost surely, we have*

$$\mathbb{P}\{(E_{n,\nu} - E_{\nu})u_{\cdot} > t\} \leq \exp\left(-\frac{1}{2} \frac{\tilde{n}t^2}{\sigma^2 + tb}\right).$$

Building on Proposition 5.2, we derive the maximal inequality over the set of functions \mathcal{U} . For $u_{\cdot} \in \mathcal{U}$, define

$$\begin{aligned} \sigma_{\nu}^2(u_{\cdot}) &= \sum_{s \in [M]} w_s \mathbb{V}_{\nu_s}[u_s(Z^{(s)})], \\ b_{\nu}(u_{\cdot}) &= \inf_b \left\{ b : \max_{s \in [M]} |u_s(Z^{(s)}) - E_{\nu_s}u_s| \leq b \text{ a.s.} \right\}. \end{aligned}$$

It is worth noting that both σ_{ν_i} and b_{ν_i} satisfy the triangle inequality. As complexity measures for \mathcal{U} , we introduce Dudley integral-type metrics, defined for $\delta > 0$ as

$$H_{\sigma, \nu_i}(\delta; \mathcal{U}) = \int_0^\delta \sqrt{\ln(N(\epsilon, \mathcal{U}, \sigma_{\nu_i}))} d\epsilon,$$

$$H_{b, \nu_i}(\delta; \mathcal{U}) = \int_0^\delta \ln(N(\epsilon, \mathcal{U}, b_{\nu_i})) d\epsilon.$$

This leads to the following maximal inequality:

Proposition 5.3. *Given a fixed $u^* \in \mathcal{U}$, $\sigma > 0$, and $b > 0$, define*

$$\mathcal{U}(\sigma, b; u^*) = \{u_i \in \mathcal{U} : \sigma_{\nu_i}(u_i - u^*) \leq \sigma, b_{\nu_i}(u_i - u^*) \leq b\}.$$

Then, there exists a universal constant $C > 0$ such that for all $t > 0$,

$$\mathbb{P}_{\nu_i^n} \left\{ \sup_{u_i \in \mathcal{U}(\sigma, b; u^*)} \sqrt{\tilde{n}}(E_{n, \nu_i} - E_{\nu_i})(u_i - u^*) \right. \\ \left. > C \left(H_{\sigma, \nu_i}(\sigma; \mathcal{U}) + \frac{1}{\sqrt{\tilde{n}}} H_{b, \nu_i}(b; \mathcal{U}) + t \right) \right\} \\ \leq \exp \left(- \frac{\sqrt{\tilde{n}} t^2}{\sqrt{\tilde{n}} \sigma^2 + t b} \right).$$

Analyses for Dudley integral-type metrics To utilize [Proposition 5.3](#), we establish upper bounds on $H_{\sigma, \nu_i}(\delta; \mathcal{U})$ and $H_{b, \nu_i}(\delta; \mathcal{U})$. In the context of analyzing $\vartheta_{n, \cdot}$, we provide upper bounds for $\mathcal{U} = \{u_{\vartheta_{n, \cdot}}^* : \nu_i \in \mathcal{Q}\}$, where \mathcal{Q} is the set of probability measures with complexity (α, β) for some $\alpha > 0$ and $\beta \geq 0$, as specified in [Definition 3.3](#).

Lemma 5.4. *Suppose that $\vartheta_{n, \cdot}^*$ satisfies the Poincaré-type inequality in [Equation \(4\)](#) for all $\nu_i \in \mathcal{Q}$, and that the complexity of \mathcal{Q} is (α, β) for some $\alpha > 0$ and $\beta \geq 0$, as in [Definition 3.3](#). Then, for all $j \geq 0$ and all $\sigma \in (0, \bar{\epsilon}]$ and $\sigma' > 0$,*

$$H_{\sigma, \nu_i}(C_P \sigma; \{u_{\vartheta_{n, \cdot}}^* - u_{\vartheta_{n, \cdot}}^* : \vartheta_i \in \Theta_j, d_{\nu_i}(\vartheta_i, \vartheta_{n, \cdot}^*) \leq \sigma'\}) \\ \leq C_P \sqrt{C' 2^{\beta j/2} (1 \wedge \sigma \wedge \sigma')} \sqrt{1 + \ln(1/(1 \wedge \sigma \wedge \sigma'))},$$

where C_P is the constant in [Equation \(4\)](#), and $\bar{\epsilon}$ and C' are the constants in [Definition 3.3](#).

Lemma 5.5. *Suppose that $\vartheta_{n, \cdot}^* \in \mathcal{M}_L$ for some $L > 1$ and all $\nu_i \in \mathcal{Q}$. Then, there exists a constant $C_b > 0$, possibly depending on M , such that for all $b > 0$,*

$$H_{b, \nu_i}(b; \{u_i - u_{\vartheta_{n, \cdot}}^* : u_i \in \mathcal{U}, b_{\nu_i}(u_i - u_{\vartheta_{n, \cdot}}^*) \leq b\}) \\ \leq C_b \sqrt{b}.$$

Relationship between potentials and transport maps

To apply existing analyses of sieved M -estimators to our setting, we establish the relationship between the potential $u_{\vartheta, \cdot}$ and the transport map ϑ, \cdot . Specifically, we prove the following:

Proposition 5.6. *Let ν_i be probability measures indexed by $[M]$ such that $\vartheta_{n, \cdot}^* \in \mathcal{M}_L^M$ for some $L > 1$. Then, for all $\vartheta_i \in \mathcal{M}_L^M$ such that $\sum_{s \in [M]} w_s \vartheta_s^{-1}(z) = z$ for all $z \in \Omega$,*

$$\frac{1}{2L} d_{\nu_i}^2(\vartheta_i, \vartheta_{n, \cdot}^*) \leq E_{\nu_i}(u_{\vartheta_i} - u_{\vartheta_{n, \cdot}}^*) \leq \frac{L}{2} d_{\nu_i}^2(\vartheta_i, \vartheta_{n, \cdot}^*).$$

Proof Sketch of Theorem 5.1 The application of [Proposition 5.3](#), together with [Lemma 5.4](#) and [Lemma 5.5](#), demonstrates that for all $k \geq 1$ and $t \geq 1$,

$$\mathbb{P}_{\nu_i^n} \{kt\tau^2 \leq d_{\nu_i}^2(\vartheta_{n, \cdot}, \vartheta_{j, \cdot}^*) \leq (k+1)t\tau^2\} \leq \exp(-kt),$$

where $\tau^2 = O((\frac{\tilde{n}}{\ln(\tilde{n})})^{-\alpha/(\alpha+\beta)})$, and $\vartheta_{j, \cdot}^*$ denotes the minimizer of $C(u_{\vartheta, \cdot}, \nu_i)$ over $\vartheta_i \in \Theta_j$. Subsequently, employing the peeling argument in conjunction with [Proposition 5.6](#) leads to the desired results.

6. Lower Bound

To establish our lower bound, we develop a technique based on reducing the fair regression estimation problem to a conventional regression estimation problem. Specifically, let $\bar{f}_{n, \cdot}^*$ be the optimal fair regression algorithm satisfying

$$\sup_{\mu_i \in \mathcal{P}} \mathbb{E}_{\mu_i} [d_{\mu_{X, \cdot}}^2(\bar{f}_{n, \cdot}^*, \bar{f}_{\mu, \cdot}^*)] = \bar{\mathcal{E}}_n(\mathcal{P}), \quad (6)$$

We demonstrate that we can construct a conventional regression algorithm using $\bar{f}_{n, \cdot}^*$. The error of this conventional regression algorithm is bounded below by $\mathcal{E}_n(\mathcal{P}_s)$, which consequently provides a lower bound on $\bar{\mathcal{E}}_n(\mathcal{P})$ in terms of $\mathcal{E}_n(\mathcal{P}_s)$.

Consider the scenario where distributions μ_1, \dots, μ_M are identical, and $\bar{f}_{n, \cdot}^*$ is constructed using samples comprising n_s i.i.d. points from μ_s . Under these conditions, we can construct a regressor for $f_{\mu, 1}^*$ as $f_n = \sum_{s \in [M]} w_s \bar{f}_{n, s}^*$. The error of this regressor provides a lower bound on [Equation \(6\)](#) as follows:

Theorem 6.1. *Under the conditions stated in [Theorem 3.4](#), we have*

$$\mathcal{E}_n(\mathcal{P}_1) \leq \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \mathbb{E}_{\mu_i} [d_{\mu_{X, 1}}^2(f_n, f_{\mu, 1}^*)] \\ \leq \sup_{\mu_i \in \mathcal{P}} \mathbb{E}_{\mu_i} [d_{\mu_{X, \cdot}}^2(\bar{f}_{n, \cdot}^*, \bar{f}_{\mu, \cdot}^*)].$$

This result directly establishes a lower bound on $\bar{\mathcal{E}}_n(\mathcal{P})$ in [Theorem 3.4](#).

7. Related Work: Optimal Transport Map Estimation

The problem of optimal transport map estimation in the Wasserstein distance has been extensively studied (Hütter & Rigollet, 2021; Divol et al., 2024; Rigollet & Stromme, 2025; Pooladian et al., 2023; Pooladian & Niles-Weed, 2024; DEB et al., 2021; del Barrio et al., 2023; Manole et al., 2024). The objective of this estimation problem is to estimate the transport map in $W_2(\mu, \mu')$ between two distributions μ and μ' , given samples from both distributions. Several approaches have been proposed: Manole et al. (2024); DEB et al. (2021); Rigollet & Stromme (2025) utilize plug-in estimators, where they first estimate the joint distribution of $(Z, \vartheta(Z))$ for $Z \sim \mu$ and subsequently construct a transport map by minimizing the expected error between Z and $\vartheta(Z)$. Alternative estimators proposed by Hütter & Rigollet (2021); Divol et al. (2024); del Barrio et al. (2023); Pooladian et al. (2023); Pooladian & Niles-Weed (2024) employ potential minimization techniques. While these studies provide convergence rate analyses for their estimators, their methods cannot be directly applied to optimal transport map estimation in the Wasserstein barycenter problem, as a sample from the barycenter distribution are not observable.

Several researchers have developed methods specifically for optimal transport maps estimation in the Wasserstein barycenter problem, although these methods lack accompanying convergence rate analyses. Korotin et al. (2020) proposed an estimator based on potential minimization, as detailed in Section 2.3. Fan et al. (2021) introduced an estimator utilizing minimax optimization. Korotin et al. (2022) developed an iterative algorithm grounded in the fixed-point theorem established by Álvarez Esteban et al. (2016). Kim et al. (2025) proposed an approach employing a primal-dual coordinate gradient algorithm to estimate the transport maps. Visentin & Cheridito (2025) developed an estimator for the transport maps leveraging the conditional normalizing flow, also demonstrating the application of their estimator to the fair regression problem. While empirical evaluations have shown that these methods achieve low estimation errors, theoretical analyses of their convergence rates remain an open problem.

8. Conclusion

We have presented a minimal optimal fair regression algorithm based on post-processing. Our algorithm achieves minimax optimality across various scenarios by leveraging the extensive body of research on minimax optimal conventional regression. Our analysis demonstrates that practitioners can focus their efforts on improving conventional regression methods, which can then be effectively adapted for fair regression applications.

Acknowledgements

This work was partly supported by JSPS KAKENHI Grant Numbers JP23K13011.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here. Rather, this paper aims at addressing a societal issue of unfairness that may arise in conventional learning algorithms.

References

- Agueh, M. and Carlier, G. Barycenters in the Wasserstein Space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011. ISSN 0036-1410. doi: 10.1137/100805741.
- Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias, 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Chen, W., Klochkov, Y., and Liu, Y. Post-hoc bias scoring is optimal for fair classification. In *The Twelfth International Conference on Learning Representations*, 2023.
- Chuang, C.-Y. and Mroueh, Y. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*, 2020.
- Chzhen, E. and Schreuder, N. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022. ISSN 0090-5364. doi: 10.1214/22-AOS2198.
- Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12739–12750, 2019.
- Chzhen, E., Denis, C., Hebiri, M., and Pontil, M. Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331. Curran Associates, Inc., 2020.
- Cotter, A., Jiang, H., Gupta, M., Wang, S., Narayan, T., Mobility Seongnam-si, K., and Korea Karthik Sridharan, S. Optimization with Non-Differentiable Constraints with Applications to Fairness, Recall, Churn, and Other Goals. *Journal of Machine Learning Research*, 20:1–59, 2019.

- Crockford, K. How is Face Recognition Surveillance Technology Racist?, 2020. URL <https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist>.
- Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women, 2018. URL <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>.
- DEB, N., Ghosal, P., and Sen, B. Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29736–29753. Curran Associates, Inc., 2021.
- del Barrio, E., Sanz, A. G., Loubes, J.-M., and Niles-Weed, J. An Improved Central Limit Theorem and Fast Convergence Rates for Entropic Transportation Costs. *SIAM Journal on Mathematics of Data Science*, 5(3):639–669, 2023. doi: 10.1137/22M149260X.
- Divol, V., Niles-Weed, J., and Pooladian, A.-A. Optimal transport map estimation in general function spaces. 2024. doi: 10.48550/arXiv.2212.03722.
- Donoho, D. L. and Johnstone, I. M. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3): 879–921, 1998. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1024691081.
- Du, M., Mukherjee, S., Wang, G., Tang, R., Awadallah, A., and Hu, X. Fairness via Representation Neutralization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12091–12103. Curran Associates, Inc., 2021.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In Goldwasser, S. (ed.), *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, Cambridge, MA, USA, 2012. ACM.
- Fan, J., Taghvaei, A., and Chen, Y. Scalable Computations of Wasserstein Barycenter via Input Convex Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1571–1581. PMLR, 2021.
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Cao, L., Zhang, C., Joachims, T., Webb, G. I., Margineantu, D. D., and Williams, G. (eds.), *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, Sydney, NSW, Australia, 2015. ACM. doi: 10.1145/2783258.2783311.
- Fukuchi, K. and Sakuma, J. Demographic parity constrained minimax optimal regression under linear model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8653–8689, 2023.
- Giné, E. and Nickl, R. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2015. ISBN 978-1-107-04316-9. doi: 10.1017/CBO9781107337862.
- Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, Barcelona, Spain, 2016. Curran Associates, Inc.
- Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2): 1166–1194, 2021. ISSN 0090-5364, 2168-8966. doi: 10.1214/20-AOS1997.
- Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein Fair Classification. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115, pp. 862–872. PMLR, 2020.
- Jovanović, N., Balunovic, M., Dimitrov, D. I., and Vechev, M. FARE: Provably Fair Representation Learning with Practical Certificates. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 15401–15420. PMLR, 2023.
- Khalili, M. M., Zhang, X., and Abroshan, M. Loss Balancing for Fair Supervised Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 16271–16290. PMLR, 2023.
- Kim, K., Yao, R., Zhu, C., and Chen, X. Optimal Transport Barycenter via Nonconvex-Concave Minimax Optimization. In *Proceedings of the 42st International Conference on Machine Learning*, 2025.
- Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, Leibniz International Proceedings in Informatics (LIPIcs), pp. 43:1–43:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. doi: 10.4230/LIPIcs.ITCS.2017.43.

- Korotin, A., Li, L., Solomon, J., and Burnaev, E. Continuous Wasserstein-2 Barycenter Estimation without Minimax Optimization. In *International Conference on Learning Representations*, 2020.
- Korotin, A., Egiazarian, V., Li, L., and Burnaev, E. Wasserstein Iterative Networks for Barycenter Estimation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15672–15686, 2022.
- Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. *The Annals of Statistics*, 52(3):966–998, 2024. ISSN 0090-5364, 2168-8966. doi: 10.1214/24-AOS2379.
- Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pp. 107–118. PMLR, 2018.
- Najibi, A. Racial Discrimination in Face Recognition Technology, 2020. URL <https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/>.
- Nishimura, Y. and Suzuki, T. Minimax optimality of convolutional neural networks for infinite dimensional input-output problems and separation from kernel methods. In *The Twelfth International Conference on Learning Representations*, 2023.
- Park, J. and Muandet, K. Towards Empirical Process Theory for Vector-Valued Functions: Metric Entropy of Smooth Function Classes. In *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, pp. 1216–1260. PMLR, 2023.
- Pedreshi, D., Ruggieri, S., and Turini, F. Discrimination-aware data mining. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, pp. 560–568, 2008. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401959.
- Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. 2024. doi: 10.48550/arXiv.2109.12004.
- Pooladian, A.-A., Divol, V., and Niles-Weed, J. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 28128–28150. PMLR, 2023.
- Rigollet, P. and Stromme, A. J. On the sample complexity of entropic optimal transport. *The Annals of Statistics*, 53(1), 2025. ISSN 0090-5364. doi: 10.1214/24-AOS2455.
- Schmidt-Hieber, J. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020. ISSN 0090-5364, 2168-8966. doi: 10.1214/19-AOS1875.
- Schreuder, N. and Chzhen, E. Classification with abstention but without disparities. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pp. 1227–1236. PMLR, 2021.
- Suzuki, T. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2018.
- Suzuki, T. and Nitanda, A. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 3609–3621. Curran Associates, Inc., 2021.
- Van Der Vaart, A. W. and Wellner, J. A. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer International Publishing, Cham, 2023. ISBN 978-3-031-29038-1 978-3-031-29040-4. doi: 10.1007/978-3-031-29040-4.
- Vigdor, N. Apple Card Investigated After Gender Discrimination Complaints. *The New York Times*, 2019. ISSN 0362-4331. URL <https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html>.
- Visentin, G. and Cheridito, P. Computing Optimal Transport Maps and Wasserstein Barycenters Using Conditional Normalizing Flows. In *Proceedings of the 42st International Conference on Machine Learning*, 2025.
- Xian, R., Yin, L., and Zhao, H. Fair and Optimal Classification via Post-Processing. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 37977–38012. PMLR, 2023.
- Xu, S. and Strohmer, T. Fair Data Representation for Machine Learning at the Pareto Frontier. *Journal of Machine Learning Research*, 24(331):1–63, 2023. ISSN 1533-7928.
- Zeng, X., Cheng, G., and Dobriban, E. Minimax Optimal Fair Classification with Bounded Demographic Disparity. 2024. doi: 10.48550/arXiv.2403.18216.
- Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating Unwanted Biases with Adversarial Learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18*, pp. 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 978-1-4503-6012-8. doi: 10.1145/3278721.3278779.

Zhao, H. and Gordon, G. Inherent Tradeoffs in Learning Fair Representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Álvarez Esteban, P. C., del Barrio, E., Cuesta-Albertos, J. A., and Matrán, C. A fixed-point approach to barycenters in Wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. ISSN 0022-247X. doi: 10.1016/j.jmaa.2016.04.045.

A. Missing Proofs

A.1. Error Analysis for General Sieved M -Estimator

In this subsection, we present an error bound for general sieved M -estimators. Let E and E_n denote the expectation and empirical process indexed by \mathcal{U} , respectively, where n is the sample size. Let Θ be a parameter space and $\Theta' \subseteq \Theta$ a sieved subset. For a family $u_\theta \in \mathcal{U}$ parameterized by $\theta \in \Theta$, the sieved M -estimator is defined as

$$\theta_n = \arg \min_{\theta \in \Theta'} E_n u_\theta.$$

Let $\theta_0 \in \Theta$ denote the ideal parameter such that $E u_{\theta_0} = \inf_{\theta \in \Theta} E u_\theta$, and let $\theta'_0 \in \Theta'$ be the ideal parameter within the sieved set, i.e., $E u_{\theta'_0} = \inf_{\theta' \in \Theta'} E u_{\theta'}$. The estimation error of θ_n is measured by $d(\theta_n, \theta_0)$, where d is a distance function on Θ .

Within this framework, we derive an error bound for sieved M -estimators under the following assumptions on the processes E and E_n .

Assumption A.1. There exist constants $K_{\text{up}}, K_{\text{low}} > 0$ such that for any $\theta \in \Theta$,

$$K_{\text{low}} E(u_\theta - u_{\theta_0}) \leq d^2(\theta, \theta_0) \leq K_{\text{up}} E(u_\theta - u_{\theta_0}).$$

Assumption A.2. Let $\gamma \in (1, 2)$ be a constant, and let $a_{0,n}, a_{1,n}, a_{2,n}, b_n > 0$ be sequences of positive numbers associated with n . Suppose $H : \mathbb{R} \rightarrow \mathbb{R}$ is a non-decreasing function such that for all $t > 0$,

$$\mathbb{P}_{\nu_n} \left\{ \sup_{\theta \in \Theta' : d(\theta, \theta') \leq \sigma} (E_n - E)(u_\theta - u_{\theta'}) > a_{0,n} H(\sigma) + a_{1,n} + a_{2,n} t \right\} \leq \exp \left(-\frac{t^2}{\sigma^2 + b_n t} \right),$$

and $\sigma \rightarrow H(\sigma)/\sigma^\gamma$ is non-increasing for $\sigma > 0$.

The following theorem provides the error bound for sieved M -estimators.

Theorem A.3. Suppose that [Assumption A.1](#) and [Assumption A.2](#) hold. Define

$$\tau_n = \sqrt{4K_{\text{up}}} \left(\sqrt{2E(u_{\theta'_0} - u_{\theta_0})} + \sqrt{a_{1,n}} + \sqrt{2h_n} \right) + \sqrt{4a_{2,n}^2 + a_{2,n}b_n}, \quad (7)$$

where h_n is a sequence such that $a_{0,n} H(\sqrt{8K_{\text{up}}h_n}) \leq h_n$ for all n . Then, for all $t \geq 1$,

$$\mathbb{P}_{\nu_n} \{ d^2(\theta_n, \theta_0) \geq t(2\tau_n^2 + 2d^2(\theta'_0, \theta_0)) \} \leq 2e^{-t}.$$

Furthermore,

$$\mathbb{E}_{\nu_n} [d^2(\theta_n, \theta_0)] \leq 4 \ln(2) \tau_n^2 + 2d^2(\theta'_0, \theta_0).$$

By establishing [Theorem A.3](#), one can obtain an error bound for a sieved M -estimator by verifying [Assumption A.1](#) and [Assumption A.2](#).

A.1.1. PROOF OF [THEOREM A.3](#)

To establish [Theorem A.3](#), we present the following supporting lemma.

Lemma A.4. Assume there exists $\tau > 0$ such that for all $k \geq 1$ and $t \geq 1$,

$$\mathbb{P}_{\nu_n} \{ kt\tau^2 \leq d^2(\hat{\theta}, \theta'_0) \leq (k+1)t\tau^2 \} \leq \exp(-kt). \quad (8)$$

Then, for all $t \geq 1$,

$$\mathbb{P}_{\nu_n} \{ d^2(\hat{\theta}, \theta_0) \geq t(2\tau^2 + 2d^2(\theta'_0, \theta_0)) \} \leq 2e^{-t}.$$

Furthermore,

$$\mathbb{E}_{\nu_n} [d^2(\hat{\theta}, \theta_0)] \leq 4 \ln(2) \tau^2 + 2d^2(\theta'_0, \theta_0).$$

Proof of Lemma A.4. Let $t > 0$ and $\tau > 0$. Define

$$\Theta_k(t) = \{\theta \in \Theta', kt\tau^2 \leq d^2(\theta, \theta'_0) \leq (k+1)t\tau^2\}.$$

By employing the peeling argument,

$$\mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta'_0) \geq t\tau^2 \right\} \leq \sum_{k=1}^{\infty} \mathbb{P}_{\nu^n} \left\{ \hat{\theta} \in \Theta_k(t) \right\}. \quad (9)$$

By assumption, for all $k \geq 1$,

$$\mathbb{P}_{\nu^n} \left\{ \hat{\theta} \in \Theta_k(t) \right\} \leq \exp(-kt). \quad (10)$$

Combining Equations (9) and (10) results in

$$\mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta'_0) \geq t\tau^2 \right\} \leq \sum_{k=1}^{\infty} \exp(-kt) = \frac{e^{-t}}{1 - e^{-t}} \leq 2e^{-t},$$

where the last inequality holds for $t \geq \ln(2)$. By the triangle inequality,

$$d^2(\hat{\theta}, \theta_0) \leq 2d^2(\hat{\theta}, \theta'_0) + 2d^2(\theta'_0, \theta_0). \quad (11)$$

Therefore, for $t \geq \ln(2)$,

$$\mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta_0) \geq 2t\tau^2 + 2d^2(\theta'_0, \theta_0) \right\} \leq 2e^{-t}.$$

For $t \geq 1$, since $2d^2(\theta'_0, \theta_0) \leq 2td^2(\theta'_0, \theta_0)$, it follows that

$$\mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta_0) \geq t(2\tau^2 + 2d^2(\theta'_0, \theta_0)) \right\} \leq 2e^{-t}.$$

For the expectation bound, for any $t_0 > 0$,

$$\begin{aligned} \mathbb{E}_{\nu^n} \left[d^2(\hat{\theta}, \theta'_0) \right] &= \int_0^{\infty} \mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta'_0) \geq x \right\} dx \\ &\leq \left(t_0 + \int_{t_0}^{\infty} \mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta'_0) \geq x\tau^2 \right\} dx \right) \tau^2 \\ &= \left(t_0 + \int_{t_0}^{\infty} \sum_{k=1}^{\infty} \mathbb{P}_{\nu^n} \left\{ \hat{\theta} \in \Theta_k(x) \right\} dx \right) \tau^2 \\ &\leq \left(t_0 + \int_{t_0}^{\infty} \sum_{k=1}^{\infty} e^{-kx} dx \right) \tau^2 \\ &= \left(t_0 + \sum_{k=1}^{\infty} \frac{e^{-kt_0}}{k} \right) \tau^2 \\ &= (t_0 - \ln(1 - e^{-t_0})) \tau^2. \end{aligned}$$

Choosing $t_0 = \ln(2)$ yields

$$\mathbb{E}_{\nu^n} \left[d^2(\hat{\theta}, \theta'_0) \right] \leq 2\ln(2)\tau^2.$$

Again, by Equation (11),

$$\mathbb{E}_{\nu^n} \left[d^2(\hat{\theta}, \theta_0) \right] \leq 4\ln(2)\tau^2 + 2d^2(\theta'_0, \theta_0).$$

□

The proof of [Theorem A.3](#) is established by confirming the condition of [Equation \(8\)](#), thereby applying [Lemma A.4](#).

Proof of Theorem A.3. By assumption and the triangle inequality, for $\theta \in \Theta'$,

$$\begin{aligned} d^2(\hat{\theta}, \theta'_0) &\leq 2d^2(\hat{\theta}, \theta_0) + 2d^2(\theta'_0, \theta_0) \\ &\leq 2K_{\text{up}}E(u_{\hat{\theta}} - u_{\theta_0}) + 4K_{\text{up}}E(u_{\theta'_0} - u_{\theta_0}) \\ &= 2K_{\text{up}}(E - \hat{E})(u_{\hat{\theta}} - u_{\theta'_0}) + 2K_{\text{up}}\hat{E}(u_{\hat{\theta}} - u_{\theta'_0}) + 4K_{\text{up}}E(u_{\theta'_0} - u_{\theta_0}) \\ &\leq 2K_{\text{up}}(E - \hat{E})(u_{\hat{\theta}} - u_{\theta'_0}) + 4K_{\text{up}}E(u_{\theta'_0} - u_{\theta_0}), \end{aligned}$$

where the last inequality uses the fact that $\hat{\theta}$ minimizes $\hat{E}u_{\theta}$ over $\theta \in \Theta'$. For notational convenience, let $U_0 = E(u_{\theta'_0} - u_{\theta_0})$. Then, the left-hand side of [Equation \(8\)](#) can be bounded as

$$\begin{aligned} &\mathbb{P}_{\nu^n} \left\{ kt\tau^2 \leq d^2(\hat{\theta}, \theta'_0) \leq (k+1)t\tau^2 \right\} \\ &\leq \mathbb{P}_{\nu^n} \left\{ d^2(\hat{\theta}, \theta'_0) \leq (k+1)t\tau^2, (E - \hat{E})(u_{\hat{\theta}} - u_{\theta'_0}) \geq \frac{kt\tau^2}{2K_{\text{up}}} - 2U_0 \right\} \\ &\leq \mathbb{P}_{\nu^n} \left\{ \sup_{\theta \in \Theta': d^2(\theta, \theta'_0) \leq (k+1)t\tau^2} (E - \hat{E})(u_{\hat{\theta}} - u_{\theta'_0}) \geq \frac{kt\tau^2}{2K_{\text{up}}} - 2U_0 \right\}. \end{aligned} \quad (12)$$

By the concentration inequality assumption and [Equation \(12\)](#), letting

$$\gamma_{k,t}(\tau) = \frac{\tau}{2K_{\text{up}}a_{2,n}} - \frac{2U_0}{a_{2,n}kt\tau} - \frac{a_{1,n}}{a_{2,n}kt\tau} - \frac{a_{0,n}}{a_{2,n}kt\tau} H(\sqrt{(k+1)t}\tau),$$

the condition in [Equation \(8\)](#) holds if

$$\frac{\gamma_{k,t}^2(\tau)}{(1 + \frac{1}{k}) + \frac{b_n}{\tau} \gamma_{k,t}(\tau)} \geq 1, \quad (13)$$

provided $\gamma_{k,t}(\tau) > 0$.

We now seek τ such that [Equation \(13\)](#) and $\gamma_{k,t}(\tau) > 0$ are satisfied. For $c, b > 0$, the function $x \mapsto \frac{x^2}{c+bx}$ is non-decreasing for $x \geq 0$. Thus, [Equation \(13\)](#) holds if a lower bound $\gamma_{k,t}(\tau) \geq \bar{\gamma}(\tau)$ satisfies

$$\frac{\bar{\gamma}^2(\tau)}{(1 + \frac{1}{k}) + \frac{b_n}{\tau} \bar{\gamma}(\tau)} \geq 1.$$

Since $\sigma \mapsto H(\sigma)/\sigma^\gamma$ is non-increasing, $H(a\sigma)/(a\sigma)^\gamma \leq H(\sigma)/\sigma^\gamma$ for all $a \geq 1$ and $\tau > 0$. Therefore, for $k \geq 1$ and $t \geq 1$,

$$\begin{aligned} \gamma_{k,t}(\tau) &\geq \frac{\tau}{2K_{\text{up}}a_{2,n}} - \frac{2U_0}{a_{2,n}kt\tau} - \frac{a_{1,n}}{a_{2,n}kt\tau} - \frac{a_{0,n}((k+1)t)^{\frac{\gamma}{2}}}{a_{2,n}kt\tau} H(\tau) \\ &\geq \frac{\tau}{2K_{\text{up}}a_{2,n}} - \frac{2U_0}{a_{2,n}\tau} - \frac{a_{1,n}}{a_{2,n}\tau} - \frac{2a_{0,n}}{a_{2,n}\tau} H(\tau) := \bar{\gamma}(\tau). \end{aligned}$$

We can rewrite $\bar{\gamma}(\tau)$ as

$$\bar{\gamma}(\tau) = \tau^{\gamma-1} \left(\frac{\tau^{2-\gamma}}{2K_{\text{up}}a_{2,n}} - \frac{2U_0}{a_{2,n}\tau^\gamma} - \frac{a_{1,n}}{a_{2,n}\tau^\gamma} - \frac{2a_{0,n}}{a_{2,n}\tau^\gamma} H(\tau) \right).$$

Since $\sigma \mapsto H(\sigma)/\sigma^\gamma$ is non-increasing, $\bar{\gamma}(\tau)$ is a product of two non-decreasing functions in τ , and thus is non-decreasing in τ when these functions are non-negative. Let $\tau^* > 0$ satisfy

$$\frac{\tau^*}{2K_{\text{up}}} - \frac{2U_0}{\tau^*} - \frac{a_{1,n}}{\tau^*} - \frac{2a_{0,n}}{\tau^*} H(\tau^*) \geq 0. \quad (14)$$

Then, for $\tau \geq \tau^*$, $\bar{\gamma}(\tau)$ is non-decreasing in τ , and for any $x > 0$, $\bar{\gamma}(\tau^* + 2K_{\text{up}}a_{2,n}x) \geq x$. Noting that $\bar{\gamma}(\tau)/\tau \leq \frac{1}{2K_{\text{up}}a_{2,n}}$, Equation (13) holds if $\bar{\gamma}^2(\tau) \geq 2 + \frac{b_n}{2K_{\text{up}}a_{2,n}}$. Therefore, for τ^* satisfying Equation (14), Equation (13) holds with

$$\tau = \tau^* + \sqrt{8K_{\text{up}}^2 a_{2,n}^2 + 2K_{\text{up}}a_{2,n}b_n}.$$

It remains to derive τ^* satisfying Equation (14). Since $\sigma \mapsto H(\sigma)/\sigma^\gamma$ is non-increasing, for any $\tau' \in (0, \tau^*]$,

$$\frac{H(\tau^*)}{\tau^*} \leq \frac{\tau^* H(\tau')}{\tau'^2}.$$

Define

$$\tau^* = \sqrt{4K_{\text{up}}} \left(\sqrt{2U_0} + \sqrt{a_{1,n}} + \sqrt{2h_n} \right),$$

where h_n is defined in the statement. Then,

$$\begin{aligned} & \frac{\tau^*}{2K_{\text{up}}} - \frac{2U_0}{\tau^*} - \frac{a_{1,n}}{\tau^*} - \frac{2a_{0,n}}{\tau^*} H(\tau^*) \\ & \geq \sqrt{\frac{4U_0}{2K_{\text{up}}}} + \sqrt{\frac{2a_{1,n}}{2K_{\text{up}}}} + \sqrt{\frac{4h_n}{2K_{\text{up}}}} - \frac{2U_0}{\sqrt{4K_{\text{up}}}\sqrt{2U_0}} - \frac{a_{1,n}}{\sqrt{4K_{\text{up}}}\sqrt{a_{1,n}}} \\ & \quad - \frac{2a_{0,n}H(\sqrt{8K_{\text{up}}h_n})}{8K_{\text{up}}h_n} \sqrt{4K_{\text{up}}} \left(\sqrt{2U_0} + \sqrt{a_{1,n}} + \sqrt{2h_n} \right) \\ & \geq \left(\sqrt{2} - \frac{2}{\sqrt{2}} \right) \sqrt{\frac{2U_0}{2K_{\text{up}}}} + \left(\sqrt{2} - \frac{2}{\sqrt{2}} \right) \sqrt{\frac{a_{1,n}}{2K_{\text{up}}}} + \left(\sqrt{2} - \frac{1}{\sqrt{2}} \right) \sqrt{\frac{2h_n}{2K_{\text{up}}}} \geq 0, \end{aligned}$$

where we use the assumption of $a_{0,n}H(\sqrt{8K_{\text{up}}h_n}) \leq h_n$ to derive the last inequality. The resulting τ then coincides with τ_n as defined in Equation (7), verifying Equation (8) with τ_n . Application of Lemma A.4 completes the proof. \square

A.2. Proofs for Section 5

A.2.1. PROOF OF THEOREM 5.1

Proof of Theorem 5.1. In the context of Appendix A.1, we identify $E = E_{\nu}$, $E_n = E_{n,\nu}$, $u_\theta = u_{\vartheta,:}$, and $d = d_{\nu,:}$. Furthermore, we set $\theta_0 = \vartheta_{\nu,:}^*$, $\theta_n = \vartheta_{n,:}$, and $\theta'_0 = \vartheta_{j,:}^*$, where $\vartheta_{j,:}^*$ denotes the minimizer of $C(u_{\vartheta,:}, \nu)$ over $\vartheta_j \in \Theta_j$. To apply Theorem A.3, it is necessary to verify Assumption A.1 and Assumption A.2.

By Proposition 5.6, the processes $E = E_{\nu}$ and $E_n = E_{n,\nu}$ satisfy Assumption A.1 with $K_{\text{low}} = \frac{1}{2L}$ and $K_{\text{up}} = \frac{L}{2}$. By the Poincaré-type inequality assumption, if $\vartheta_j \in \Theta_j$ satisfies $d_{\nu}(\vartheta_j, \vartheta_{\nu,:}^*) \leq \sigma$, then $E_{\nu}(u_{\vartheta_j} - u_{\vartheta_{\nu,:}^*}) \leq C_P \sigma$. Since Ω is bounded, for $\vartheta_j \in \Theta_j \subseteq \mathcal{M}_L^M$, there exists a constant $b > 0$ such that $b_{\nu}(u_{\vartheta_j} - u_{\vartheta_{\nu,:}^*}) \leq b$, which follows from the smoothness of $u_{\vartheta,s}$ and $u_{\vartheta_{\nu,:}^*,s}$. Therefore, by applying Proposition 5.3 together with Lemma 5.4 and Lemma 5.5, these processes satisfy Assumption A.2 with

$$\begin{aligned} a_{0,n} &= \frac{CC_P \sqrt{C'} 2^{\beta j/2}}{\sqrt{\tilde{n}}} \\ a_{1,n} &= \frac{CC_b \sqrt{b}}{\tilde{n}} \\ a_{2,n} &= \frac{C}{\sqrt{\tilde{n}}} \\ b_n &= \frac{b}{\sqrt{\tilde{n}}} \\ H(\sigma) &= (1 \wedge \sigma) \sqrt{1 + \ln(1/(1 \wedge \sigma))}, \end{aligned}$$

where C is the constant in [Proposition 5.3](#). With the choice of j as stated, we have

$$a_{0,n} \leq CC_P \sqrt{C'} \frac{\tilde{n}^{-\alpha/2(\alpha+\beta)}}{\ln^{\beta/2(\alpha+\beta)}(\tilde{n})}.$$

Set $h_n = c(\ln(\tilde{n})/\tilde{n})^{\alpha/(\alpha+\beta)}$ for some $c > 0$ such that $4Lh_n \leq 1$. Then,

$$\frac{a_{0,n}H(\sqrt{4Lh_n})}{h_n} \leq CC_P \sqrt{\frac{4C'L}{c}} \ln^{-1/2}(\tilde{n}) \sqrt{1 + \frac{\alpha}{2(\alpha+\beta)} \ln(\tilde{n}) - \frac{1}{2} \ln\left(4Lc \ln^{\frac{\alpha}{\alpha+\beta}}(\tilde{n})\right)}. \quad (15)$$

The right-hand side of [Equation \(15\)](#) is monotonically decreasing in c and \tilde{n} . Hence, for sufficiently large c independent of \tilde{n} , we have $\frac{a_{0,n}H(\sqrt{4Lh_n})}{h_n} \leq 1$. Thus, by applying [Theorem A.3](#), we obtain the error bound with

$$\tau_n = \sqrt{2L} \left(\sqrt{2E(u_{\vartheta_j^*,:} - u_{\vartheta_{\nu^*}^*,:})} + \sqrt{\frac{CC_b}{\tilde{n}}} + \sqrt{2c \left(\frac{\ln(\tilde{n})}{\tilde{n}}\right)^{\alpha/(\alpha+\beta)}} + \sqrt{\frac{4C^2}{\tilde{n}} + \frac{CC_b}{\tilde{n}}} \right).$$

From [Proposition 5.6](#) and [Definition 3.3](#), it follows that

$$\begin{aligned} E_{\nu^*}(u_{\vartheta_j^*,:} - u_{\vartheta_{\nu^*}^*,:}) &= \inf_{\vartheta} E_{\nu^*}(u_{\vartheta,:} - u_{\vartheta_{\nu^*}^*,:}) \\ &\leq 2L \inf_{\vartheta} d_{\nu^*}^2(\vartheta, \vartheta_{\nu^*}^*) \leq 2LC2^{-\alpha j}, \end{aligned}$$

and

$$d_{\nu^*}^2(\vartheta_j^*, \vartheta_{\nu^*}^*) \leq \frac{L}{2} E_{\nu^*}(u_{\vartheta_j^*,:} - u_{\vartheta_{\nu^*}^*,:}) \leq 2L^2C2^{-\alpha j}.$$

With the choice of j as stated, we have

$$2^{-\alpha j} \leq \frac{1}{2^\alpha} \left(\frac{\tilde{n}}{\ln(\tilde{n})} \right)^{-\alpha/(\alpha+\beta)}.$$

Therefore,

$$\begin{aligned} \tau_n^2 &= O\left(\left(\frac{\tilde{n}}{\ln(\tilde{n})}\right)^{-\alpha/(\alpha+\beta)}\right) \\ d_{\nu^*}^2(\vartheta_j^*, \vartheta_{\nu^*}^*) &= O\left(\left(\frac{\tilde{n}}{\ln(\tilde{n})}\right)^{-\alpha/(\alpha+\beta)}\right), \end{aligned}$$

which establishes the claim. □

A.2.2. PROOF OF [LEMMA 5.4](#) AND [LEMMA 5.5](#)

Here, we present the proofs of [Lemma 5.4](#) and [Lemma 5.5](#).

Proof of Lemma 5.4. By the Poincaré-type inequality,

$$N(\epsilon, \{u_{\vartheta,:} - u_{\vartheta_{\nu^*}^*,:} : \vartheta \in \Theta_j\}, \sigma_{\nu^*}) \leq N\left(\frac{\epsilon}{C_P}, \Theta_j, d_{\nu^*}\right).$$

According to [Definition 3.3](#), we have

$$\ln N(\epsilon, \{\vartheta : \vartheta \in \Theta_j, d_{\nu^*}(\vartheta, \vartheta_{\nu^*}^*) \leq \sigma'\}, d_{\nu^*}) \leq C'2^{\beta j} \ln_+(1/\epsilon),$$

for $\epsilon \leq \sigma' \wedge \bar{\epsilon}$, and the log-covering number is zero if $\epsilon > \sigma'$. Therefore,

$$\begin{aligned} & H_{\sigma, \nu} (C_P \sigma; \{u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot} : \vartheta \in \Theta_j, d_{\nu}(\vartheta, \vartheta^*) \leq \sigma'\}) \\ &= \int_0^{C_P \sigma} \sqrt{\ln N(\epsilon, \{u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot} : \vartheta \in \Theta_j, d_{\nu}(\vartheta, \vartheta^*) \leq \sigma'\}, \sigma_{\nu})} d\epsilon \\ &\leq \int_0^{C_P(\sigma \wedge \sigma')} \sqrt{C' 2^{\beta j} \ln_+(C_P/\epsilon)} d\epsilon \\ &\leq C_P \sqrt{C' 2^{\beta j/2}} \int_0^{\sigma \wedge \sigma'} \sqrt{\ln_+(1/\epsilon)} d\epsilon. \end{aligned}$$

By the Cauchy-Schwarz inequality, for any $z \in (0, 1)$,

$$\begin{aligned} \int_0^z \sqrt{\ln(1/\epsilon)} d\epsilon &\leq \sqrt{\int_0^z 1 d\epsilon \int_0^z \ln(1/\epsilon) d\epsilon} \\ &= \sqrt{z(z \ln(1/z) + z)} \leq z \sqrt{1 + \ln(1/z)}. \end{aligned}$$

Thus,

$$H_{\sigma, \nu} (C_P \sigma; \{u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot} : \vartheta \in \Theta_j, d_{\nu}(\vartheta, \vartheta^*) \leq \sigma'\}) \leq C_P \sqrt{C' 2^{\beta j/2}} (1 \wedge \sigma \wedge \sigma') \sqrt{1 + \ln(1/(1 \wedge \sigma \wedge \sigma'))}.$$

□

Proof of Lemma 5.5. Let $\|\cdot\|_\infty$ denote the ∞ -norm over Ω . For any \mathcal{U} , $N(\epsilon, \mathcal{U}, b_{\nu}) \leq N_\infty(\epsilon, \mathcal{U}, \|\cdot\|_\infty)$, since $\nu \in \mathcal{Q}$ is supported only on Ω . For any $\vartheta \in \mathcal{M}_L^M$, $u_{\vartheta, s}$ belongs to the Hölder class with smoothness 2. Interpreting $u_{\vartheta, \cdot}$ as a vector-valued function, it follows from [Park & Muandet \(2023\)](#) that there exists a constant $C > 0$, possibly depending on M , such that $\ln N(\epsilon, \{u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot} : \vartheta \in \mathcal{M}_L^M\}, \|\cdot\|_\infty) \leq C\epsilon^{-1/2}$. Therefore,

$$H_{b, \nu}(b; \{u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot} : \vartheta \in \mathcal{M}_L^M, b_{\nu}(u_{\vartheta, \cdot} - u_{\vartheta^*, \cdot}) \leq b\}) \leq \int_0^b C\epsilon^{-1/2} d\epsilon = 2C\sqrt{b},$$

which establishes the claim. □

A.2.3. PROOF OF PROPOSITION 5.2

Proof of Proposition 5.2. We derive a high probability upper bound through bounding the cumulant generating function. Given a random variable X on a probability space $(\mathcal{Z}, \mathfrak{F}, \nu)$, define the cumulant generating function of X as for $\lambda \in \mathbb{R}$,

$$\kappa(\lambda; X) = \ln(\mathbb{E}_\nu[\exp(\lambda X)]).$$

The Chernoff bound implies that for $\lambda \geq 0$,

$$\mathbb{P}_\nu\{X \geq t\} \leq \exp(-\lambda t + \kappa(\lambda; X)). \quad (16)$$

Hence, an upper bound on the cumulant generating function yields the corresponding high probability upper bound.

Let $Z_1^{(s)}, \dots, Z_{n_s}^{(s)}$ be n_s i.i.d. copies of $Z^{(s)}$. Let $\sigma_s^2 = \mathbb{V}_{\nu_s}[u_s(Z^{(s)})]$, and let b_s be the minimum real such that $u_s(Z^{(s)}) - \mathbb{E}_{\nu_s}[u_s(Z^{(s)})] \leq b_s$ almost surely. For $\lambda > 0$, we derive an upper bound on the cumulant generating function of $(E_{n, \nu} - E_{\nu})u_{\cdot}$. By the independence among $Z^{(\cdot)}$, we have

$$\kappa(\lambda; (E_{n, \nu} - E_{\nu})u_{\cdot}) = \sum_{s \in [M]} n_s \ln \left(\mathbb{E}_{\nu_s} \left[\exp \left(\lambda \frac{w_s}{n_s} \left(u_s(Z_1^{(s)}) - \int u_s(z) \nu_s(dz) \right) \right) \right] \right).$$

Noting that the Bernstein's condition is satisfied for a bounded random variable, we obtain

$$\ln \left(\mathbb{E}_{\nu_s} \left[\exp \left(\lambda \frac{w_s}{n_s} \left(u_s(Z_1^{(s)}) - \int u_s(z) \nu_s(dz) \right) \right) \right] \right) \leq \frac{\left| \lambda \frac{w_s}{n_s} \right|^2 \sigma_s^2}{2 \left(1 - b_s \left| \lambda \frac{w_s}{n_s} \right| \right)},$$

provided that $\left| \lambda \frac{w_s}{n_s} \right| \leq \frac{1}{b_s}$. Hence,

$$\kappa(\lambda; (E_{n,\nu_\cdot} - E_{\nu_\cdot})u_\cdot) \leq \sum_{s \in [M]} n_s \frac{\lambda^2 \frac{w_s^2}{n_s^2} \sigma_s^2}{2 \left(1 - b_s \left| \lambda \frac{w_s}{n_s} \right| \right)},$$

as long as $\left| \lambda \frac{w_s}{n_s} \right| \leq \frac{1}{b_s}$ for all $s \in [M]$.

Define $\bar{\sigma}^2 = \sum_{s \in [M]} \frac{w_s^2 \sigma_s^2}{n_s} \leq \frac{\sigma^2}{\bar{n}}$ and $\bar{b} = \max_{s \in [M]} \frac{w_s b_s}{n_s} \leq \frac{b}{\bar{n}}$. Set $\lambda = \frac{t}{\bar{\sigma}^2 + t\bar{b}}$, which satisfies $0 \leq \lambda \frac{w_s}{n_s} \leq \frac{1}{b_s}$ for all $s \in [M]$. Then, we have

$$\begin{aligned} & \kappa(\lambda; (E_{n,\nu_\cdot} - E_{\nu_\cdot})u_\cdot) - \lambda t \\ & \leq \sum_{s \in [M]} n_s \frac{t^2 \frac{w_s^2}{n_s^2} \sigma_s^2}{2(\bar{\sigma}^2 + t\bar{b})^2 \left(1 - b_s \left| \lambda \frac{w_s}{n_s} \right| \right)} - \frac{t^2}{\bar{\sigma}^2 + t\bar{b}} \\ & = \sum_{s \in [M]} n_s \frac{t^2 \frac{w_s^2}{n_s^2} \sigma_s^2}{2(\bar{\sigma}^2 + t\bar{b}) \left(\bar{\sigma}^2 + t\bar{b} - t \frac{w_s b_s}{n_s} \right)} - \frac{t^2}{\bar{\sigma}^2 + t\bar{b}} \\ & \leq \sum_{s \in [M]} n_s \frac{t^2 \frac{w_s^2}{n_s^2} \sigma_s^2}{2(\bar{\sigma}^2 + t\bar{b}) \bar{\sigma}^2} - \frac{t^2}{\bar{\sigma}^2 + t\bar{b}} = -\frac{t^2}{2(\bar{\sigma}^2 + t\bar{b})} \leq -\frac{\tilde{n} t^2}{2(\sigma^2 + tb)}. \end{aligned} \tag{17}$$

The claim follows by applying the Chernoff bound in Equation (16) with the upper bound derived in Equation (17). \square

A.2.4. PROOF OF PROPOSITION 5.3

Proof of Proposition 5.3. The desired claim is obtained by applying Van Der Vaart & Wellner (Theorem 2.2.19 and Lemma 2.2.15 2023) with Proposition 5.2 and

$$d_2(u_\cdot, u'_\cdot) = \sigma_{\nu_\cdot}^2(u_\cdot - u'_\cdot) \quad d_1(u_\cdot, u'_\cdot) = \frac{1}{\sqrt{\tilde{n}}} b_{\nu_\cdot}(u_\cdot - u'_\cdot).$$

\square

A.2.5. PROOF OF PROPOSITION 5.6

Proof of Proposition 5.6. Let $\vartheta_\cdot \in \mathcal{M}_L^M$ be arbitrary. We begin by expressing

$$\begin{aligned} & E_{\nu_\cdot} u_{\vartheta_\cdot} \\ & = \sum_{s \in [M]} w_s \int u_{\vartheta,s}(z) \nu_s(dz) \\ & = \sum_{s \in [M]} w_s \int \left(u_{\vartheta,s}(z) + u_{\vartheta,s}^\dagger(\vartheta_{\nu,s}^*(z)) \right) \nu_s(dz) - \sum_{s \in [M]} w_s \int u_{\vartheta,s}^\dagger(\vartheta_{\nu,s}^*(z)) \nu_s(dz) \\ & = \sum_{s \in [M]} w_s \int \left(u_{\vartheta,s}(z) + u_{\vartheta,s}^\dagger(\vartheta_{\nu,s}^*(z)) \right) \nu_s(dz) - \int \sum_{s \in [M]} w_s u_{\vartheta,s}^\dagger(z) \nu(dz), \end{aligned}$$

where the last equality follows from the change of variables $\nu_s = \vartheta_{\nu,s}^{*-1} \# \nu$. By the assumption that $\sum_{s \in [M]} w_s \vartheta_s^{-1}(z) = z$ for all $z \in \Omega$, we obtain

$$\int \sum_{s \in [M]} w_s u_{\vartheta,s}^\dagger(z) \nu(dz) = \int \sum_{s \in [M]} w_s \int_0^z \vartheta_s^{-1}(x) dx \nu(dz)$$

$$\begin{aligned}
 &= \int \int_0^z \sum_{s \in [M]} w_s \vartheta_s^{-1}(x) dx \nu(dz) \\
 &= \int \int_0^z x dx \nu(dz) = \frac{1}{2} \int z^2 \nu(dz).
 \end{aligned}$$

Therefore, we have

$$\begin{aligned}
 &\sum_{s \in [M]} w_s \int u_{\vartheta, s}(z) \nu_s(dz) \\
 &= \sum_{s \in [M]} w_s \int \left(u_{\vartheta, s}(z) + u_{\vartheta, s}^\dagger(\vartheta_{\nu, s}^*(z)) \right) \nu_s(dz) - \frac{1}{2} \int z^2 \nu(dz).
 \end{aligned} \tag{18}$$

Next, for any $\vartheta'_s \in \mathcal{M}_L^M$, we derive upper and lower bounds for $u_{\vartheta, s}(z) + u_{\vartheta, s}^\dagger(\vartheta'_s(z))$ by employing the approach of [Hütter & Rigollet \(2021\)](#). Since $\vartheta_{\cdot, \cdot} \in \mathcal{M}_L^M$, the function $u_{\vartheta, \cdot}$ is L -smooth and $1/L$ -strongly convex. Thus, for all $z, x \in \mathbb{R}$, it holds that

$$\frac{1}{2L} |z - x|^2 \leq u_{\vartheta, s}(z) - u_{\vartheta, s}(x) - \vartheta_s(x)(z - x) \leq \frac{L}{2} |z - x|^2.$$

We first establish the lower bound. For $z, x \in \mathbb{R}$, define

$$q_x(z) = u_{\vartheta, s}(x) + \vartheta_s(x)(z - x) + \frac{L}{2} |z - x|^2.$$

The convex conjugate of q_x at $\vartheta'_s(x)$ is given by

$$q_x^\dagger(\vartheta'_s(x)) = -u_{\vartheta, s}(x) + x\vartheta'_s(x) + \frac{1}{2L} (\vartheta'_s(x) - \vartheta_s(x))^2.$$

Since for convex functions f, g , $f \leq g$ implies $f^\dagger \geq g^\dagger$, it follows that for all $z \in \mathbb{R}$ and any $\vartheta_{\cdot, \cdot}, \vartheta'_{\cdot, \cdot} \in \mathcal{M}_L^M$,

$$u_{\vartheta, s}(z) + u_{\vartheta, s}^\dagger(\vartheta'_s(z)) \geq z\vartheta'_s(z) + \frac{1}{2L} (\vartheta'_s(z) - \vartheta_s(z))^2. \tag{19}$$

The upper bound is obtained analogously by considering

$$q_x(z) = u_{\vartheta, s}(x) + \vartheta_s(x)(z - x) + \frac{1}{2L} |z - x|^2,$$

leading to

$$u_{\vartheta, s}(z) + u_{\vartheta, s}^\dagger(\vartheta'_s(z)) \leq z\vartheta'_s(z) + \frac{L}{2} (\vartheta'_s(z) - \vartheta_s(z))^2. \tag{20}$$

We now combine Equations (18) to (20) to establish the claim. From [Equation \(18\)](#), we have

$$\begin{aligned}
 &E_{\nu_{\cdot}}(u_{\vartheta, \cdot} - u_{\vartheta_{\nu}^*, \cdot}) \\
 &= \sum_{s \in [M]} w_s \int \left(u_{\vartheta, s}(z) + u_{\vartheta, s}^\dagger(\vartheta_{\nu, s}^*(z)) - u_{\vartheta_{\nu}^*, s}(z) + u_{\vartheta_{\nu}^*, s}^\dagger(\vartheta_{\nu, s}^*(z)) \right) \nu_s(dz).
 \end{aligned} \tag{21}$$

Observe that $Du_{\vartheta', s}(z) = \vartheta'_s(z)$, and the equality case of the Young-Fenchel inequality gives

$$u_{\vartheta', s}(z) + u_{\vartheta', s}^\dagger(\vartheta'_s(z)) = z\vartheta'_s(z). \tag{22}$$

Applying Equations (19) and (22) with $\vartheta'_s = \vartheta_{\nu, s}^*$ in [Equation \(21\)](#), we obtain

$$E_{\nu_{\cdot}}(u_{\vartheta, \cdot} - u_{\vartheta_{\nu}^*, \cdot})$$

$$\begin{aligned}
 &\geq \sum_{s \in [M]} w_s \int \frac{1}{2L} (\vartheta_s(z) - \vartheta_{\nu,s}^*(z))^2 \nu_s(dz) \\
 &= \frac{1}{2L} \sum_{s \in [M]} w_s d_{\nu_s}^2(\vartheta_s, \vartheta_{\nu,s}^*) = \frac{1}{2L} d_{\nu_\cdot}^2(\vartheta_\cdot, \vartheta_{\nu,\cdot}^*)
 \end{aligned}$$

Similarly, by using Equation (20) in place of Equation (19), we have

$$E_{\nu_\cdot}(u_{\vartheta,\cdot} - u_{\vartheta_{\nu,\cdot}^*,\cdot}) \leq \frac{L}{2} d_{\nu_\cdot}^2(\vartheta_\cdot, \vartheta_{\nu,\cdot}^*)$$

□

A.3. Proofs for Section 4

A.4. Proof of Corollary 4.2

Proof of Corollary 4.2. According to Theorem 5.1 and Proposition 4.5, there exists a decreasing sequence r_n and a constant $C > 0$ such that, for all $t \geq 1$,

$$\mathbb{P}_{\mu_\cdot^n} \left\{ \inf_{\mu} \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \mu) > C t r_n \right\} \leq e^{-t}.$$

By choosing $t = r_n^{-1/2}$, it follows that

$$\mathbb{P}_{\mu_\cdot^n} \left\{ \inf_{\mu} \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \mu) > o(1) \right\} \leq o(1).$$

Therefore, $\inf_{\mu} \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \mu)$ converges to 0 in probability. In the case where $\inf_{\mu} \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \mu) = 0$, it holds that $\bar{f}_{n,1} \# \mu_{X,1} = \dots = \bar{f}_{n,M} \# \mu_{X,M}$, which satisfies (strict) demographic parity. Consequently, $f_{n,\cdot}$ is consistently fair. The error upper bound can be established directly by combining Theorem 5.1 and Proposition 4.4. □

A.4.1. PROOF OF PROPOSITION 4.4

Proof of Proposition 4.4. We begin by applying the triangle inequality, which yields

$$d_{\mu_{X,\cdot}}(\bar{f}_{n,\cdot}, \bar{f}_{\mu,\cdot}^*) \leq d_{\mu_{X,\cdot}}(\bar{f}_{n,\cdot}, \vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{n,\cdot}) + d_{\mu_{X,\cdot}}(\vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{n,\cdot}, \vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{\mu,\cdot}^*) + d_{\mu_{X,\cdot}}(\vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{\mu,\cdot}^*, \bar{f}_{\mu,\cdot}^*). \quad (23)$$

The first term in (23) represents the estimation error of the optimal transport map:

$$\begin{aligned}
 d_{\mu_{X,\cdot}}^2(\bar{f}_{n,\cdot}, \vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{n,\cdot}) &= \sum_{s \in [M]} w_s \int \left((\vartheta_{n,s} \circ f_{n,s})(x) - (\vartheta_{\mu_{\hat{f},s}}^* \circ f_{n,s})(x) \right)^2 \mu_{X,s}(dx) \\
 &= \sum_{s \in [M]} w_s \int \left(\vartheta_{n,s}(z) - \vartheta_{\mu_{\hat{f},s}}^*(z) \right)^2 \mu_{\hat{f},s}(dz) = d_{\mu_{\hat{f},\cdot}}^2(\vartheta_{n,\cdot}, \vartheta_{\mu_{\hat{f},\cdot}}^*).
 \end{aligned}$$

For the second term in (23), we establish the L -Lipschitz continuity of $\vartheta_{\mu_{\hat{f},\cdot}}^*$. By the first assumption in Assumption 3.1, $f_{n,\cdot} \in \mathcal{F}$ implies $f_{n,\cdot} \# \mu_{X,\cdot} \in \mathcal{P}_{\mathcal{F}}$, which ensures that $\vartheta_{\mu_{\hat{f},\cdot}}^* \in \Theta$. The L -Lipschitz continuity of $\vartheta_{\mu_{\hat{f},\cdot}}^*$ then follows from Assumption 3.2. Therefore, we can bound the second term as

$$\mathbb{E}_{\mu_\cdot^{2n}} \left[d_{\mu_{X,\cdot}}^2(\vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{n,\cdot}, \vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{\mu,\cdot}^*) \right] \leq L^2 \mathbb{E}_{\mu_\cdot^n} \left[d_{\mu_{X,\cdot}}^2(f_{n,\cdot}, f_{\mu,\cdot}^*) \right] \leq L^2 \sum_{s \in [M]} w_s \mathcal{E}_{n_s}(\mathcal{P}_s).$$

For the third term in (23), we utilize Proposition 5.6 to obtain

$$d_{\mu_{X,\cdot}}^2(\vartheta_{\mu_{\hat{f},\cdot}}^* \circ f_{\mu,\cdot}^*, \bar{f}_{\mu,\cdot}^*) = d_{\mu_{\hat{f},\cdot}}^2(\vartheta_{\mu_{\hat{f},\cdot}}^*, \vartheta_{\mu_{\hat{f},\cdot}}^*)$$

$$\begin{aligned}
 &\leq 2LE_{\mu_{f,:}} \left(u_{\vartheta_{\mu_{\hat{f},:}}^*} - u_{\vartheta_{\mu_{f,:}}^*} \right) \\
 &\leq 2L \left(E_{\mu_{f,:}} - E_{\mu_{\hat{f},:}} \right) \left(u_{\vartheta_{\mu_{\hat{f},:}}^*} - u_{\vartheta_{\mu_{f,:}}^*} \right) \\
 &\leq 2L \left| \left(E_{\mu_{f,:}} - E_{\mu_{\hat{f},:}} \right) u_{\vartheta_{\mu_{\hat{f},:}}^*} \right| + 2L \left| \left(E_{\mu_{f,:}} - E_{\mu_{\hat{f},:}} \right) u_{\vartheta_{\mu_{f,:}}^*} \right|
 \end{aligned}$$

Let φ_s be mappings such that $\varphi_s \# \mu_{\hat{f},s} = \mu_{f,s}$. For $\vartheta_s = \vartheta_{\mu_{f,s}}^*$ or $\vartheta_{\mu_{\hat{f},s}}^*$, we have

$$\begin{aligned}
 \left| \left(E_{\mu_{f,s}} - E_{\mu_{\hat{f},s}} \right) u_{\vartheta,s} \right| &= \left| \int (u_{\vartheta,s}(\varphi_s(z)) - u_{\vartheta,s}(z)) \mu_{\hat{f},s}(dz) \right| \\
 &= \left| \int \int_z^{\varphi_s(z)} \vartheta_s(x) dx \mu_{\hat{f},s}(dz) \right| \\
 &\leq \int \frac{L}{2} (\varphi_s(z) - z)^2 \nu(dz) \\
 &= \frac{L}{2} W_2^2(\mu_{f,s}, \mu_{\hat{f},s}) \leq \frac{L}{2} d_{\mu_s}^2(f_{n,:}, f_{\mu_s}^*),
 \end{aligned}$$

where we use the identity $u_{\vartheta,s}(x) - u_{\vartheta,s}(y) = \int_y^x \vartheta_s(z) dz$ for all $x, y \in \mathbb{R}$, which follows from Taylor's theorem, to obtain the second equality. Taking expectations, we obtain

$$\mathbb{E}_{\mu_s^{2n}} \left[d_{\mu_{X,:}}^2 \left(\vartheta_{\mu_{f,:}}^* \circ f_{\mu_s}^*, \bar{f}_{\mu_s}^* \right) \right] \leq 2L^2 \sum_{s \in [M]} w_s \mathbb{E}_{\mu_s^{2n}} [\mathcal{E}_{n_s}(\mathcal{P}_s)] \leq 2L^2 \sum_{s \in [M]} w_s \mathcal{E}_{n_s}(\mathcal{P}_s).$$

By combining the above results, we have

$$\mathbb{E}_{\mu_s^{2n}} \left[d_{\mu_{X,:}}^2(\bar{f}_{n,:}, \bar{f}_{\mu_s}^*) \right] \leq 3\mathbb{E}_{\mu_s^{2n}} \left[d_{\mu_{\hat{f},:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) \right] + 9L^2 \sum_{s \in [M]} w_s \mathcal{E}_{n_s}(\mathcal{P}_s),$$

where Jensen's inequality is applied. Finally, due to the mutual independence between $f_{n,:}$ and $\vartheta_{\mu_{\hat{f},:}}^*$ resulting from sample splitting, we have $\mathbb{E}_{\mu_s^{2n}} \left[d_{\mu_{X,:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) \right] = \mathbb{E}_{\mu_{\hat{f},:}^{2n}} \left[d_{\mu_{\hat{f},:}}^2(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) \right]$. This completes the proof. \square

A.4.2. PROOF OF PROPOSITION 4.5

Proof of Proposition 4.5. Let $\mu_{\hat{f}}$ be the barycenter of $\mu_{\hat{f},s}$ with weight w_s . For any probability measure ν , we have

$$\begin{aligned}
 W_2(\bar{f}_{n,s} \# \mu_{X,s}, \nu) &= W_2(\vartheta_{n,s} \# \mu_{\hat{f},s}, \nu) \leq W_2(\vartheta_{n,s} \# \mu_{\hat{f},s}, \vartheta_{\mu_{\hat{f},s}}^* \# \mu_{\hat{f},s}) + W_2(\vartheta_{\mu_{\hat{f},s}}^* \# \mu_{\hat{f},s}, \nu) \\
 &\leq d_{\mu_{\hat{f},s}}(\vartheta_{n,s}, \vartheta_{\mu_{\hat{f},s}}^*) + W_2(\mu_{\hat{f}}, \nu).
 \end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
 \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \nu) &\leq \max_{s \in [M]} d_{\mu_{\hat{f},s}}(\vartheta_{n,s}, \vartheta_{\mu_{\hat{f},s}}^*) + W_2(\mu_{\hat{f}}, \nu) \\
 &\leq \sqrt{\frac{1}{Mw_{\min}}} d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) + W_2(\mu_{\hat{f}}, \nu).
 \end{aligned}$$

Consequently, it follows that

$$\inf_{\nu} \max_{s \in [M]} W_2(\bar{f}_{n,s} \# \mu_{X,s}, \nu) \leq \sqrt{\frac{1}{Mw_{\min}}} d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*) + \inf_{\nu} W_2(\mu_{\hat{f}}, \nu) = \sqrt{\frac{1}{Mw_{\min}}} d_{\mu_{\hat{f},:}}(\vartheta_{n,:}, \vartheta_{\mu_{\hat{f},:}}^*).$$

\square

A.5. Proofs for Section 6

Proof of Theorem 6.1. We begin by noting that, by definition, $\bar{f}_{n,s} \in \mathcal{F}_s$. According to the second requirement in Assumption 3.1, it holds that $\mathcal{F}_1 = \dots = \mathcal{F}_M$. Furthermore, the third requirement in Assumption 3.1 ensures that $f_n \in \mathcal{F}_1$. Therefore, f_n can be interpreted as the standard regression estimator for \mathcal{P}_1 based on a sample of size n . By the definition of the minimax error, we have

$$\sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \mathbb{E}_{\mu_n} \left[d_{\mu_{X,1}}^2(f_n, f_{\mu,1}^*) \right] \geq \mathcal{E}_n(\mathcal{P}_1).$$

Next, utilizing the convexity of $d_{\mu_1}^2$, we obtain

$$\begin{aligned} \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \mathbb{E}_{\mu_n} \left[d_{\mu_{X,1}}^2(f_n, f_{\mu,1}^*) \right] &\leq \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \sum_{s \in [M]} w_s \mathbb{E}_{\mu_n} \left[d_{\mu_{X,1}}^2(\bar{f}_{n,s}^*, f_{\mu,1}^*) \right] \\ &= \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \sum_{s \in [M]} w_s \mathbb{E}_{\mu_n} \left[d_{\mu_{X,s}}^2(\bar{f}_{n,s}^*, f_{\mu,s}^*) \right]. \end{aligned}$$

When $\mu_1 = \dots = \mu_M$, the transport map $\vartheta_{\mu,s}^*$ becomes the identity function for all s . Thus,

$$\begin{aligned} \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \sum_{s \in [M]} w_s \mathbb{E}_{\mu_n} \left[d_{\mu_{X,s}}^2(\bar{f}_{n,s}^*, f_{\mu,s}^*) \right] &= \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \sum_{s \in [M]} w_s \mathbb{E}_{\mu_n} \left[d_{\mu_{X,s}}^2(\bar{f}_{n,s}^*, \bar{f}_{\mu,s}^*) \right] \\ &= \sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \mathbb{E}_{\mu_n} \left[d_{\mu_{X,:}}^2(\bar{f}_{n,:}^*, \bar{f}_{\mu,:}^*) \right]. \end{aligned}$$

Moreover, by the second requirement in Assumption 3.1, for any $\mu \in \mathcal{P}_1$, the tuple $\underbrace{(\mu, \dots, \mu)}_{M \text{ times}}$ belongs to \mathcal{P} . Therefore,

$$\sup_{\mu_1 \in \mathcal{P}_1: \forall s \in [M], \mu_s = \mu_1} \mathbb{E}_{\mu_n} \left[d_{\mu_{X,:}}^2(\bar{f}_{n,:}^*, \bar{f}_{\mu,:}^*) \right] \leq \sup_{\mu \in \mathcal{P}} \mathbb{E}_{\mu_n} \left[d_{\mu_{X,:}}^2(\bar{f}_{n,:}^*, \bar{f}_{\mu,:}^*) \right].$$

□