Fair and Optimal Classification via Transports to Wasserstein-Barycenter

Ruicheng Xian  Lang Yin  Han Zhao
University of Illinois Urbana-Champaign
\{rxian2,langyin2,hanzhao\}@illinois.edu

Abstract

Fairness in automated decision-making systems has gained increasing attention as their applications expand to real-world high-stakes domains. To facilitate the design of fair ML systems, it is essential to understand the potential trade-offs between fairness and predictive power, and the construction of the optimal predictor under a given fairness constraint. In this paper, for general classification problems under the group fairness criterion of demographic parity (DP), we precisely characterize the trade-off between DP and classification accuracy, referred to as the minimum cost of fairness. Our insight comes from the key observation that finding the optimal fair classifier is equivalent to solving a Wasserstein-barycenter problem under \(\ell_1\)-norm restricted to the vertices of the probability simplex. Inspired by our characterization, we provide a construction of an optimal fair classifier achieving this minimum cost via the composition of the Bayes regressor and optimal transports from its output distributions to the barycenter. Our construction naturally leads to an algorithm for post-processing any pre-trained predictor to satisfy DP fairness, complemented with finite sample guarantees. Experiments on real-world datasets verify and demonstrate the effectiveness of our approaches.

1 Introduction

The increasing applications of machine learning for automated decision-making in high-stakes domains, including criminal justice, healthcare, finance, and online advertising, have prompted studies on aspects of fairness of machine learning models (Berk et al., 2021). For instance, models trained on past data may propagate or exacerbate bias against historically underrepresented or disadvantaged demographic groups (Barocas and Selbst, 2016). To address the potential bias, works in algorithmic fairness formulate concrete mathematical criteria that captures various notions of fairness that the model shall satisfy. Different types of algorithms designed to produce fair models have been proposed, including data pre-processing (Calmon et al., 2017; Song et al., 2019), constrained optimization (Zafar et al., 2017; Agarwal et al., 2019), and post-processing (Hardt et al., 2016; Zhao and Gordon, 2022) under learning paradigms of supervised learning (Hardt et al., 2016; Zhao et al., 2020; Madras et al., 2018), unsupervised learning (Chierichetti et al., 2017; Li et al., 2020; Backurs et al., 2019), ranking (Zehlike et al., 2017), and sequential decision making (Joseph et al., 2016; Gillen et al., 2018; Joseph et al., 2017; Chi et al., 2022).

Broadly speaking, there are two categories of definitions for algorithmic fairness, namely individual fairness (Dwork et al., 2012) and group fairness (Verma and Rubin, 2018; Zemel et al., 2013). As its name suggests, a predictive model is said to satisfy individual fairness if it treats
similar individuals similarly. The main challenge of adopting this notion of individual fairness in practice lies in the specification of concrete metrics to measure the similarity of individuals, which must be application and context dependent. On the other hand, group fairness is often defined to be the equality of certain statistical quantities, e.g., true positive/negative rates (Hardt et al., 2016), accuracy (Buolamwini and Gebru, 2018) etc., among different subgroups in the overall population. Among various notions of algorithmic fairness, one of the most-studied group fairness criterion is demographic parity (DP), which states that the output of the model should be statistically independent of the demographic group membership of the input (Calders et al., 2009), perhaps due to its close relation to the “80% rule” in disparate impact legislature (Feldman et al., 2015).

In general, fairness criteria impose limits on the predictive power of models under consideration because they are viewed as constraints on the hypothesis space. Take demographic parity as an example, the model cannot be simultaneously fair and accurate if the target variable of interest is strongly correlated with the demographic group membership. Hence for each criterion and problem instance, there should be a minimum cost of fairness: the difference between the maximum unconstrained model performance and that under the fairness constraint. Characterizing such costs is not only of theoretical interest, but would also better inform practitioners about the trade-offs imposed by the fairness constraint, which can further inspire better designs of fair ML systems.

For demographic parity, recent work has established the cost of fairness for regression problems in terms of mean squared error under the setting of attribute-aware predictors (Le Gouic et al., 2020; Chzhen et al., 2020). Zhao and Gordon (2022) have also studied this problem for binary classification problems under the realizable setting (i.e., noiseless setting with zero Bayes error rate) when the number of demographic groups is two. However, it remains an open problem for classification in the general setting on what is the minimum cost of fairness and how to achieve it. In this paper, we focus on the above problem under the most general setting of classification, i.e., multiclass, multigroup and agnostic learning (non-zero Bayes error). In particular, we show that:

1. The minimum cost of DP fairness, defined as the increase in classification error of the optimal (randomized) classifier with and without the DP constraint, is given by the optimal value of a Wasserstein-barycenter problem under \( \ell_1 \)-norm, where the barycenter is a finite distribution supported on the vertices of the probability simplex (Theorem 1).
2. This characterization reveals that the optimal fair classifier is the composition of the Bayes regressor of the one-hot labels and the optimal transports from its output distributions to the barycenter (Theorem 3). Since the supports of the distributions considered in the transportation problem are the simplex and its vertices, we refer to this as the simplex-vertex transportation problem.
3. Our theoretical insights suggest a three-step method for obtaining optimal fair classifiers: learn the Bayes regressor, find the Wasserstein-barycenter, and compute the optimal transports to the barycenter (Section 3.1). In practical scenarios where the Bayes regressor is not available, the last two steps can be applied to any pre-trained predictors as a post-processing procedure to derive fair classifiers (Section 3.2 and Algorithm 1).
4. We further provide novel procedures for obtaining finite-sample estimates of the barycenter and the optimal transports for our proposed method above, with generalization guarantees (Section 4 and Algorithm 2).

Although our main focus is theoretical, we also perform experiments on real-world benchmark datasets to demonstrate the effectiveness of our proposed post-processing algorithm (Section 5).
Table 1: Minimum costs of fairness for demographic parity.

| Problem Setting            | Minimum Cost of Fairness                     | Expression                                                                 |
|---------------------------|----------------------------------------------|-----------------------------------------------------------------------------|
| Regression                | Mean squared err., excess risk               | \( \min_{q: \text{supp}(q) \subseteq R} \sum_{a \in A} w_a \cdot W_2^2(r_a^*, q) \) (Wasserstein-barycenter under \( \ell_2 \)) |
| Classification, realizable| Classification err., excess = min. risk      | \( \min_{q: \text{supp}(q) \subseteq \{e_1, \ldots, e_k\}} \sum_{a \in A} w_a \cdot d_{TV}(p_a^*, q) \) (TV-barycenter) |
| Classification            | Classification err., minimum risk            | \( \min_{q: \text{supp}(q) \subseteq \{e_1, \ldots, e_k\}} \sum_{a \in A} w_a \cdot \frac{1}{2} W_1(r_a^*, q) \) (Wasserstein-barycenter under \( \ell_1 \)) |

1.1 Related Work

The cost of DP fairness is studied in a number of recent work under different problem settings. For regression, where the risk is the mean squared error, Le Gouic et al. (2020) and Chzhen et al. (2020) concurrently established that the minimum cost of DP fairness, or the excess risk achieved by the optimal regressor satisfying DP fairness w.r.t. the unconstrained, attribute-aware Bayes regressor, \( f^*(x, a) := E[Y \mid X = x, A = a] \), is given by the optimal value of a Wasserstein-barycenter problem under 2-norm (first row of Table 1) where \( r_a^* := f^*\sharp(\mu_X \times \{a\}) \) is the output distribution of the attribute-aware Bayes regressor \( f^* \) conditioned on group \( a \) (supported on \( \mathbb{R} \)), the \( q \)'s are distributions supported on \( \mathbb{R} \) with finite second moments (the minimizer is called the barycenter), and the \( w_a \)'s are nonnegative weights assigned to each group.

We show in Section 3 that for classification, the minimum classification error achieved by the optimal randomized DP fair classifier, or the minimum risk, is given by the optimal value of a Wasserstein-barycenter problem under \( \ell_1 \)-norm (third row of Table 1); excess risk is minimum risk less the Bayes error of the problem at hand. Here, \( f^* \) is the Bayes regressor of the labels \( Y \) in one-hot representation, whereby the \( r_a^* \)'s are supported on the \((k - 1)\)-simplex (\( k \) is the number of classes).

This expression is very similar to the one for regression problems, but there are two main differences. First, the metric here is the \( W_1 \) distance under \( \ell_1 \)-norm, reflecting the fact that the classification error is the average total variation (TV) distance between the true class probabilities and the one-hot class labels over all instances. In contrast, for the regression setting, the metric used in \( W_2^2(\cdot, \cdot) \) is the \( \ell_2 \) norm instead, which is induced from the use of mean squared error. Second, the candidate barycenters \( q \) are restricted to distributions supported on vertices of the simplex, corresponding to the one-hot labels.

The connection between Wasserstein-barycenters and classification under DP is explored in a number of prior and concurrent works (Jiang et al., 2020, Zhao and Gordon, 2022, Denis et al., 2022, Gaucher et al., 2022). Jiang et al. (2020) consider the post-processing of pre-trained predictors leveraging Wasserstein-barycenters, but not the minimum cost of DP fairness. The latter works establish the minimum cost, but only for the special cases of realizable setting (i.e., zero Bayes error), binary sensitive attribute (\(|A| = 2\)), or binary class labels (\( k = 2 \)), respectively. Our result above, established for the general classification setting of multiclass, multigroup and agnostic learning (non-zero Bayes error), can recover existing ones. For instance, Zhao and Gordon (2022).
establish that under the realizable setting the minimum cost of DP fairness is related to a TV-barycenter problem (second row of Table 1) where \( p^*_a \) is the distribution of class labels on group \( a \), \( p^*_a(c_i) = \mathbb{P}_\mu(Y = c_i \mid A = a) \), and also supported on the vertices; this in fact coincides with \( r^*_a \), which is a special case of Wasserstein-barycenter (Proposition 7).

2 Preliminaries

Let the \( k \)-class classification problem be defined by a joint distribution \( \mu \) of input \( X \), class label \( Y \), and demographic group membership \( A \) (a.k.a. sensitive attribute), supported on \( \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \). The labels \( Y \) are in one-hot representation, \( \mathcal{Y} = \{e_1, \cdots, e_k\} \) where \( e_i := (0, \cdots, 0, 1, 0, \cdots, 0) \) with the 1 on the \( i \)-th coordinate, and a finite number of demographic groups is assumed, \( \mathcal{A} = [m] := \{1, \cdots, m\} \).

The goal of fair classification is to find a fair and optimal attribute-aware classifier \( h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y} \) on the classification problem, satisfying the designated fairness criteria and achieving the minimum group-balanced classification error rate:

\[
\sum_{a \in [m]} \text{Err}_a(h) := \sum_{a \in [m]} \mathbb{P}(h(X, a) \neq Y \mid A = a) \\
= \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \mathbb{P}(h(x, a) \neq y) \, d\mathbb{P}_{\mu_a}(X = x, Y = y),
\]

where we defined \( \mu_a \) to be the distribution of \( \mu \) conditioned on \( A = a \). Note that since the classifier is allowed access to \( A \), the problem is equivalent to finding \( m \) classifiers \( h_1, \cdots, h_m : \mathcal{X} \to \mathcal{Y} \) on the classification problems given by \( \mu_1, \cdots, \mu_m \), each over one demographic group, such that their combination via \( h(x, a) := h_a(x) \) is fair and optimal on the original problem \( \mu \).

The present work considers the group fairness criterion of demographic parity, and we call a classifier fair as long as it satisfies DP:

**Definition 1** (Demographic Parity). A classifier \( h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y} \) satisfies demographic parity if for all \( y \in \mathcal{Y} \), and \( a, a' \in \mathcal{A} \),

\[
\mathbb{P}(h(X, a) = y \mid A = a) = \mathbb{P}(h(X, a') = y \mid A = a') \\
= \int_{x \in \mathcal{X}} \mathbb{P}(h(x, a) = y) \, d\mathbb{P}_{\mu_a}(X = x) = \int_{x \in \mathcal{X}} \mathbb{P}(h(x, a') = y) \, d\mathbb{P}_{\mu_{a'}}(X = x).
\]

When a classifier fails to satisfy DP exactly, we measure its fairness as the violation of DP by the maximum TV distance between its output distributions on any two groups, referred to as the DP gap:

\[
\text{DP Gap}(h) := \frac{1}{2} \max_{a, a' \in \mathcal{A}} \sum_{y \in \mathcal{Y}} |\mathbb{P}(h(X, a) = y \mid A = a) - \mathbb{P}(h(X, a') = y \mid A = a')|.
\]

We allow the classifier to be randomized, whose output on an input is not deterministic but follows a (conditional) distribution. The error rate of a randomized classifier and its DP fairness are evaluated w.r.t. not only the population \( \mu \) but also to the randomness of its outputs, as illustrated by the decompositions in Eqs. (1) and (2). A formal definition of randomized functions via Markov kernels is provided below. Unless explicitly stated, all functions in this paper are randomized.

---

1 Our results (Theorem 1) can be extended to arbitrarily weighted error in Appendix A.1. For clarity, the discussions in the main sections are focused on the balanced error w.l.o.g.
Definition 2 (Markov Kernel). A Markov kernel from a measurable space \((X, \mathcal{S})\) to \((Y, \mathcal{T})\) is a mapping \(K : X \times T \to [0, 1]\), such that \(K(\cdot, T)\) is \(\mathcal{S}\)-measurable \(\forall T \in \mathcal{T}\), and \(K(x, \cdot)\) is a probability measure on \((Y, \mathcal{T}) \ \forall x \in X\).

Definition 3 (Randomized Function). A randomized function \(f : X \to Y\) is defined by a Markov kernel \(K : X \times \mathcal{B}(Y) \to [0, 1]\), where \(\mathcal{B}(Y)\) denotes Borel \(\sigma\)-algebra on \(Y\), such that \(\forall x \in X, B \in \mathcal{B}(Y), \ P(f(x) \in B) = K(x, B)\).

Our presentation uses push-forward measure for notational convenience, whose definition is extended to randomized functions as follows:

Definition 4 (Push-Forward by Randomized Function). Let a measure \(p\) on \(X\) and a randomized function \(f : X \to Y\) with Markov kernel \(K\) be given. The push-forward of \(p\) under \(f\), denoted by \(f^\ast p\), is a measure on \(Y\) where \(f^\ast p(B) := \int_{x \in X} K(x, B) \, dp(x)\) for all \(B \in \mathcal{B}(Y)\).

With these definitions, Definition 1 is equivalently and compactly written as

\[ h^\sharp(\mu_a^X \times \{a\}) \overset{d}{=} h^\sharp(\mu_{a'}^X \times \{a'\}), \quad \forall a, a' \in A; \]

the equality is in distribution, and \(\mu_a^X\) denotes the marginal distribution of \(X\) under \(\mu\).

Optimal Transport and Wasserstein Distance. Classification under DP fairness constraint is closely related to optimal transport and Wasserstein distance, for which a brief overview is given below. Interested readers are referred to Villani (2003) for a more complete treatment.

Definition 5 (Coupling). Let \(p, q\) be probability measures on \(X\) and \(Y\), respectively. A coupling of \(p\) and \(q\) is a joint distribution \(\gamma\) of \(X \times Y\) s.t. \(p(x) = \int_{y \in Y} \, d\gamma(x, y)\) for all \(x \in X\), and \(q(y) = \int_{x \in X} \, d\gamma(x, y)\) for all \(y \in Y\). Let \(\Gamma(p, q)\) denote the collection of all couplings of \(p\) and \(q\).

Definition 6 (Optimal Transport). Let \(p, q\) be probability measures on \(X\) and \(Y\), respectively, and \(c : X \times Y \to [0, \infty)\) a cost function. The optimal transportation cost between \(p\) and \(q\) is

\[ \inf_{\gamma \in \Gamma(p, q)} \int_{X \times Y} c(x, y) \, d\gamma(x, y). \]

Given an optimal coupling \(\gamma^*\) that achieves the infimum in the definition above, an optimal transport from \(p\) to \(q\) can be obtained, denoted by \(T^*_{p \to q} : X \to Y\), which is the randomized function with Markov kernel \(K(x, B) = \gamma^*(x, B)/\gamma^*(x, Y)\). All optimal transports in the present work concern measures whose supports are contained in \(\Delta^k\), under the \(\ell_1\) cost of \(c(x, y) = \|x - y\|_1\).

When \(X = Y\) and are metric spaces equipped with distance function \(d\), the optimal transportation cost between \(p\) and \(q\) under \(c = d\) coincides with their Wasserstein-1 distance:

Definition 7 (Wasserstein Distance). Let \(p, q\) be probability measures on a metric space \((X, d)\). For \(r \geq 1\), the Wasserstein-\(r\) distance between \(p\) and \(q\) is

\[ W_r(p, q) := \left(\inf_{\gamma \in \Gamma(p, q)} \int_{X \times X} d(x, x')^r \, d\gamma(x, x')\right)^{1/r}. \]

3 Fair Classification as a Barycenter Problem

We relate the problem of learning fair classifiers to that of finding a Wasserstein-barycenter, and use this equivalence to derive methods for computing optimal fair classifiers (Section 3.1), as well as post-processing pre-trained predictors to satisfy DP fairness (Section 3.2).
Definitions. First, we define several quantities that appear in our results; see Fig. 1 for illustrations. Let \( \Delta_k := \{ x \in \mathbb{R}^k : x \geq 0, \| x \|_1 = 1 \} \) denote the \((k-1)\)-dimensional probability simplex. Given a classification problem \( \mu \), let \( f^*: \mathcal{X} \times \mathcal{A} \to \Delta_k \) denote the (deterministic) Bayes regressor of the one-hot class labels \( Y \): the minimum mean squared error estimator of \( Y \) given \( X \) and \( A \),

\[
f^*(x,a) = \mathbb{E}_\mu[Y \mid X = x, A = a] = \mathbb{P}_\mu(Y = e. \mid X = x, A = a),
\]

whose outputs \( f^*(x,a) \in \Delta^k \) are the probability mass of the true classes \( Y \) in vector form conditioned on each input \( x \in \text{supp}(\mu^X) \) and group \( a \). \( f^* \) is interpreted as the optimal predictor on \( \mu \). Let \( r^*_a := f^*\sharp(\mu^X \times \{ a \}) \), which is supported on the simplex and represents the distribution of the true class probabilities conditioned on inputs from group \( a \). Finally, let \( \mathcal{Q}_k \) denote the collection of distributions supported on \( \mathcal{Y} = \{ e_1, \cdots , e_k \} \), the vertices of the simplex.

Our main result states that the minimum error rate, or minimum risk, among all randomized classifiers satisfying DP fairness is given by the optimal value of a barycenter problem on the \( r^*_a \)'s under the Wasserstein-1 distance with the \( \ell_1 \) metric restricted to the vertices of the simplex; the excess risk is obtained by subtracting the Bayes error on \( \mu \), and thereby giving a characterization of the minimum cost of DP fairness:

**Theorem 1** (Minimum Cost of Fairness). Let \( \mu \) be given, and \( f^*: \mathcal{X} \times \mathcal{A} \to \Delta_k \) denote the Bayes regressor on \( \mu \), \( r^*_a := f^*\sharp(\mu^X \times \{ a \}) \), and \( \mathcal{Q}_k \) the set of distributions on \( \mathcal{Y} = \{ e_1, \cdots , e_k \} \). Then

\[
\min_{h: \text{DP}} \sum_{a \in [m]} \text{Err}_a(h) = \frac{1}{2} \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} W_1(r^*_a, q)
\]

\[
= \frac{1}{2} \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} \min_{\gamma_\mathcal{Y} \in \Gamma(r^*_a, q)} \int_{\Delta_k \times \mathcal{Y}} \| s - y \|_1 \, d\gamma_\mathcal{Y}(s, y).
\]

Remarks. Several remarks are in place. First, DP fairness requires that the distribution of output classifications to remain the same when conditioned on any group, and the barycenter \( q^* \) is the output class distribution that gives the minimum DP fair classification error.

Second, the Bayes error, or the minimum error rate without the DP constraint, is given by

\[
\sum_a \min_{h_a} \mathbb{P}(Y \neq h_a(X) \mid A = a) = \frac{1}{2} \sum_a \min_{q_a} W_1(r^*_a, q_a) \leq \frac{1}{2} \min_q \sum_a W_1(r^*_a, q),
\]

where Eq. (4) is included in the last expression for comparison. The last inequality is nonstrict even when \( Y \perp A \), meaning that the minimum cost of fairness for classification is zero on certain problem instances, i.e., the Bayes error can be achieved by an unfair classifier as well as a fair one; an example based on the nonuniqueness of the optimal decision rule is provided below. This is in contrast to regression problems, where the minimum cost of DP fairness is zero if and only if \( Y \perp A \).
Example. Consider the two-group binary classification problem $\mu$ satisfying

$$P_{\mu_1}(Y = e_1) = 1 \quad \text{and} \quad P_{\mu_2}(Y = e_1 \mid X = x) = P_{\mu_2}(Y = e_2 \mid X = x) = \frac{1}{2}, \forall x \in X.$$  

The optimal decision rule on group 1 is the constant function $x \mapsto e_1$. All rules on group 2 yield the same (therefore optimal) accuracy of 1/2 due to the randomness in $Y$, including the decision rule of $x \mapsto e_1$, which when combined with the optimal rule on group 1 satisfies DP.

Lastly, the Wasserstein-barycenter expression can be simplified under certain assumptions: when the Bayes error is zero, it reduces to the TV-barycenter problem of $\min_q \sum_a d_{TV}(p_a^*, q)$, where $p_a^*$ is the true class distribution conditioned on group $a$, $p_a^*(e_i) := P_{\mu}(Y = e_i \mid A = a)$. This equality is established by Zhao and Gordon (2022) but only for $k = m = 2$. The proof of this reduction is included in Appendix A.2.

Proof Sketch. Theorem 1 is a direct consequence of a reformulation of the classification problem when oracle access to the class probabilities of each $(x, a) \in \text{supp}(\mu^{X, A})$ is available (here they are provided by $f^*$). With this equivalence, the classification problem becomes one of finding the optimal assignment of class labels given class probabilities. Clearly the best assignment is the one that outputs the class with the highest probability, $x \mapsto e_{\arg\max_i f^*(x)_i}$, but it may not satisfy DP.

To restrict the hypothesis space to classifiers that satisfy DP, we can consider a bi-level search, where in the outer loop we fix an output distribution $q \in \mathcal{Q}_k$ that the assignment should obey when conditioned on each group, and then in the inner loop we find the best assignment with the output distribution $q$. This might not seem to be simpler than the original problem at first, but the following lemma shows that the optimal value of the inner loop is the optimal transportation cost, $W_1(r_a^*, q)$, effectively reducing the bi-level problem to a linear program:

**Lemma 2.** Let $\mu$ with only one demographic group be given, let $f^*(x) := E_{\mu}[Y \mid X = x]$ denote the Bayes regressor on $\mu$, and define $r^* := f^* \mu^X$. Fix $q \in \mathcal{Q}_k$. Then for any randomized classifier $h : X \rightarrow Y$ satisfying $h^* \mu^X \equiv q$, there exists a coupling $\gamma \in \Gamma(r^*, q)$ s.t.

$$\text{Err}(h) := P(h(X) \neq Y) = \frac{1}{2} \int_{\Delta_k \times Y} ||s - y||_1 d\gamma(s, y). \quad (5)$$

Conversely, for any $\gamma \in \Gamma(r^*, q)$, there exists a randomized classifier $h$ satisfying $h^* \mu^X \equiv q$ s.t. Eq. (5) holds.

Proof of Theorem 1. Lemma 2 shows that for fixed $q \in \mathcal{Q}_k$, the minimum error on group $a$ is $\text{Err}_a(h) = \frac{1}{2} W_1(r_a^*, q)$, so the theorem follows by summing up the error over all groups and minimizing $q \in \mathcal{Q}_k$. ■

### 3.1 Exact Computation of an Optimal Fair Classifier

Theorem 1 shows the existence of a randomized DP fair classifier with the minimum cost of fairness given by Eq. (4), but the theorem does not tell us how to obtain such a classifier. In what follows, we provide a construction for this optimal fair classifier $\bar{h}^*$ given the barycenter $q^*$ from Eq. (4) and the Bayes regressor $f^*$, which is simply the composition of $f^*$ and the simplex-vertex optimal transports $T_{r_a^* \rightarrow q^*}$ from $r_a^* := f^*(\mu_a^X \times \{a\})$ to $q^*$.
Algorithm 1: Post-process a predictor for DP fairness

1. function PostProcess($f; \mu^X$) \text{ \textsuperscript{\textdagger}} $f : \mathcal{X} \times \mathcal{A} \rightarrow \Delta_k$ is a predictor on the classification problem $\mu$, and $\mu^X$ is the marginal input distribution

2. for $a \in [m]$ do

3. \hspace{0.5cm} $r_a \leftarrow f^\sharp(\mu^X \times \{a\})$

4. \hspace{0.5cm} $q^* \leftarrow \arg \min_{q \in Q_k} \sum_{a \in [m]} W_1(r_a, q)$ \text{ \textsuperscript{\textdagger}} find the barycenter as in Eq. (4)

5. for $a \in [m]$ do

6. \hspace{0.5cm} Compute $T_{r_a \rightarrow q^*}^* : \Delta_k \rightarrow \mathcal{Y}$ \text{ \textsuperscript{\textdagger}} optimal transport from $r_a$ to $q^*$ under the $\ell_1$ cost

7. return $(x, a) \mapsto T_{r_a \rightarrow q^*}^* \circ f(x, a)$

\begin{align}
\text{Theorem 3 (Optimal Fair Classifier).} \ & \text{Fix } q \in Q_k, \text{ and let } T_{r_a \rightarrow q^*}^* : \Delta_k \rightarrow \mathcal{Y} \text{ denote the optimal transport from } r_a^* \text{ to } q \text{ under the } \ell_1 \text{ cost of } c(s, y) = \| s - y \|_1. \text{ Then the classifier } h \text{ given by}
\end{align}

$$ h(x, a) = T_{r_a \rightarrow q^*}^* \circ f^*(x, a) $$

\text{satisfies } $\sum_{a \in [m]} \text{Err}_a(h) = \frac{1}{2} \sum_{a \in [m]} W_1(r_a^*, q) \text{ and } \forall a \in [m], \bar{h}_a^*(\mu_a^X \times \{a\}) \overset{d}{=} q \text{, i.e., it is DP fair.}$

Therefore, fair and optimal classifications on $\mu$ are obtained via a three-step method. First, learn the Bayes regressor $f^*$ of the one-hot labels $Y$ from minimizing the squared loss,

$$ f^* = \min_f \mathbb{E}_{\mu}(\| f(X, A) - Y \|^2_2). $$

Compute the barycenter $q^*$ of the $r_a^*$’s, and finally the optimal transports $T_{r_a \rightarrow q^*}^*$ from the $r_a^*$’s to $q^* \in Q_k$. Then by Theorem 3, $h^*(x, a) := T_{r_a \rightarrow q^*}^* \circ f^*(x, a)$ is an optimal fair classifier. This is summarized in Algorithm 1 with the Bayes regressor $f^*$ as the argument.

In practice, however, we may not obtain $f^*$ exactly due to difficulties in representation, optimization, and (finite sample) generalization, or it could simply be too expensive to train a perfect predictor (from scratch). Most likely we will be working with suboptimal predictors $\tilde{f} \approx f^*$ (pre-trained by a vendor). Then, could we still use PostProcess to patch $\tilde{f}$ for satisfying DP fairness? Yes indeed; the derivation of fair classifiers from arbitrary pre-trained predictors is discussed next.

3.2 Approximating Optimal Fair Classifier from Pre-Trained Predictor

Let $\tilde{f} : \mathcal{X} \times \mathcal{A} \rightarrow \Delta_k$ be an arbitrary and potentially suboptimal deterministic pre-trained predictor (approximate of the Bayes regressor $f^*$). The goal of this section is to derive classifiers from $\tilde{f}$ with (randomized) post-processing maps $g_a : \Delta_k \rightarrow \mathcal{Y}$ for each $a \in [m]$, via $(x, a) \mapsto g_a \circ \tilde{f}(x, a)$, such that the derived classifier satisfies DP fairness, and ideally, achieves the minimum error rate among all fair classifiers derived from $\tilde{f}$.

Define $\tilde{r}_a := \tilde{f}^\sharp(\mu_a^X \times \{a\})$. If we pretend $\tilde{f}$ to be the Bayes regressor regardless and call PostProcess($\tilde{f}$) of Algorithm 1 (which only requires $(X, A)$ pairs without class labels), the returned classifier $\tilde{h}(x, a) := T_{\tilde{r}_a \rightarrow \tilde{q}}^* \circ \tilde{f}(x, a)$ satisfies DP fairness since $\tilde{h}_a^*(\mu_a^X \times \{a\}) \overset{d}{=} \tilde{q}$ for all $a \in [m]$, where $\tilde{q}$ denotes the barycenter of the $\tilde{r}_a$’s obtained by solving the barycenter problem on Line 4 of Algorithm 1 its error rate, however, depends on the suboptimality of $\tilde{f}$ relative to the Bayes regressor $f^*$ in $L^1$:
**Theorem 4** (Error Propagation). For any fixed $\mu$, let $f^*$ denote the Bayes regressor on $\mu$, and $\text{Err}^*$ the minimum fair error rate on $\mu$ as in Eq. (4). Let $h$ denote the DP fair classifier derived from a given deterministic predictor $\tilde{f} : \mathcal{X} \times \mathcal{A} \to \Delta_k$ via \textit{PostProcess}($\tilde{f} ; \mu^X$). Then

$$\sum_{a \in [m]} \text{Err}_a(h) - \text{Err}^* \leq \sum_{a \in [m]} \mathbb{E}_{\mu^X} [\|\tilde{f}(X,a) - f^*(X,a)\|_1].$$

However, this $\tilde{h}$ does not necessarily incur the minimum cost of fairness among all DP classifiers, nor even among those derived from $\tilde{f}$. But if the predictor $\tilde{f}$ is \textit{calibrated}, the classifier returned from \textit{PostProcess} will be optimal among all derived DP classifiers.

**Definition 8** (Calibration). Let $\mu$ with only one demographic group be given. A deterministic predictor $f : \mathcal{X} \to \Delta_k$ is \textit{calibrated} if $\mathbb{P}_\mu(Y = e_i | f(X) = s) = s_i$ for all $s \in \Delta_k$ and $i \in [k]$.

An uncalibrated predictor $\tilde{f}$ can be calibrated by applying a calibration map $u^*$, s.t. the optimal (fair) classifier derived from $(x,a) \mapsto u^*(\tilde{f}(x,a),a)$ has the same error rate as that of $\tilde{f}$, is by definition the Bayes regressor of $Y$ given $(\tilde{f}(X,A),A)$, hence it can be learned from labeled data via minimizing the squared loss between $u^*(\tilde{f}(X,A),A)$ and $Y$ (when $\tilde{f}$ is already calibrated, $u^*(s,a) = s$, i.e., an identity map). Calibration maps could also be learned using potentially more efficient methods, e.g., Guo et al. (2017) calibrates neural networks via histogram binning.

Note that finding an optimal fair post-processing map on $\tilde{f}$ is equivalent to finding an optimal fair classifier on a new classification problem with a transformed input space, namely the joint distribution of $(\tilde{f}(X,A),Y,A)$. We know by Theorem 3 that the optimal fair classifier on $\tilde{f}$ is the composition of $u^*(s,a) := \mathbb{E}_{\tilde{\mu}}[Y | X = s, A = a]$—which coincides with the optimal calibration map—and the optimal transports from the distributions of $u^*(\tilde{f}(X,A),A)$ to their barycenter; this is exactly the post-processing map applied to the calibrated predictor $(x,a) \mapsto u^*(\tilde{f}(x,a),a)$ by \textit{PostProcess} (along with $\mu^X$ as the argument), thereby the classifier it returns is optimal among all fair classifiers derived from $\tilde{f}$.

### 4 Algorithms for Finite-Sample Estimates

Algorithm 1 as discussed in Sections 3.1 and 3.2 outlines a method for learning optimal DP fair classifiers on any given classification problem $\mu$, and also for post-processing arbitrary pre-trained predictors $f$ to satisfy DP fairness using \textit{unlabeled} data. However, it assumes access to the underlying joint distribution $\mu$, which is not available to the learner in practice; instead, only finite samples drawn from $\mu$ will be available. In this section, we focus on Lines 4–6 of computing the Wasserstein-barycenter (Section 4.1) and the simplex-vertex optimal transports (Section 4.2), discuss procedures for obtaining finite-sample estimates, and analyze their generalization properties.

Our finite-sample generalization guarantees bound the DP gap (Eq. (3)) of the derived classifier on the population, as well as the error rate. For the latter, we assume that the pre-trained classifier $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ is calibrated (Definition 8), and state the suboptimality relative to the optimal DP fair classifier derived from $f$, whose error rate, by the discussions in Section 3.2 is

$$\text{Err}_f^* := \frac{1}{2} \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} W_1(r_a,q)$$

where $r_a := f^*_a(\mu^X_a \times \{a\})$. In addition, for the analyses, we assume that the i.i.d. (unlabeled) samples of $(f(X,A),A)$ used for the estimation of the barycenter and the optimal transports are
not previously involved in the optimization or tuning of \( f \):
\[
S_a = \{ s_{a,i} \}_{i=1}^n := \{ f(x_{a,i}, a) \}_{i=1}^n, \quad x_{a,i} \sim \mu_a^X \text{ or equivalently } s_{a,i} \sim r_a, \quad \forall a \in [m],
\]
and \( f \) does not depend on these \( S_a \)'s.

### 4.1 Linear Program for Wasserstein-Barycenter

We find or estimate the Wasserstein-barycenter via solving a linear program (LP). Suppose for now that the supports of the \( r_a \)'s are finite, denoted by \( \mathcal{R}_a := \text{supp}(r_a) \). If we know their true probability mass, then their barycenter \( q^* \) can be obtained from the minimizer of the following LP:

\[
\text{OPT}(r_1, \cdots, r_m) : \quad \min_{\gamma_1, \cdots, \gamma_m} \sum_{a \in [m]} \sum_{s \in \mathcal{R}_a, y \in \mathcal{Y}} \|s - y\|_1 \cdot \gamma_a(s, y)
\]
\[
\text{s.t.} \quad \sum_{y \in \mathcal{Y}} \gamma_a(s, y) = r_a(s), \quad \forall a \in [m], s \in \mathcal{R}_a,
\]
\[
\sum_{y \in \mathcal{Y}} \gamma_1(s, y) = \sum_{s \in \mathcal{R}_a} \gamma_a(s, y), \quad \forall a \in \{2, \cdots, m\}, y \in \mathcal{Y},
\]
\[
\sum_{s \in \mathcal{R}_a, y \in \mathcal{Y}} \gamma_1(s, y) = 1, \quad \gamma_a \geq 0, \quad \forall a \in [m],
\]

where \( \gamma_a \in \mathbb{R}^{[\mathcal{R}_a] \times k} \). The program has \( \left( \sum_{a \in [m]} \sum_{k|\mathcal{R}_a|} \right) \) design variables and inequality constraints, and \( \left( \sum_{a \in [m]} |\mathcal{R}_a| + k(m - 1) + 1 \right) \) equality constraints. Note that it is an implementation of the Wasserstein-barycenter problem in Eq. (4) for the finite case, therefore, the barycenter is extracted from an optimal solution \( \{ \gamma_a^* \}_{a \in [m]} \) as a vector in \( \mathbb{R}^k \) representing its pmf via

\[
q^* = \sum_{a \in \mathcal{R}_a} \gamma_a^*(s, y), \quad \text{for arbitrary } a \in [m],
\]

As will be discussed soon, we can also extract transports from the \( r_a \)'s to \( q^* \).

If the \( r_a \)'s are not finite or their probability mass are unknown, and instead we are working with finite samples \( S_a \), then we proceed with the empirical distributions \( \hat{r}_a := \frac{1}{n} \sum_{i=1}^n \delta_{s_{a,i}} \), where \( \delta \) denotes the Dirac delta function, and solve OPT(\( \hat{r}_1, \cdots, \hat{r}_m \)) to get an estimated barycenter \( \hat{q} \).

### 4.2 Extracting Transports from Minimizers of OPT

After (the estimated) Wasserstein-barycenter is obtained from solving the OPT above, we compute (or estimate) the optimal post-processing map for the predictor \( f \) towards satisfying DP fairness, which is given by the simplex-vertex optimal transport maps from the \( r_a \)'s to the barycenter. Similar to the barycenter, the transports can also be extracted from the solution to OPT: indeed, by definition it is the optimal transports when the distributions \( r_a := f^*_\#(\mu_a^X \times \{a\}) \) are finite (Section 4.2.1). But when they are not finite, extra handling is required to generalize the transports to unseen data (Sections 4.2.2 and 4.2.3).

#### 4.2.1 The Finite Case

When the \( r_a \)'s are finite and their probability mass are known, the optimal solution \( \{ \gamma_a^* \}_{a \in [m]} \) to OPT already provides the optimal transports: the randomized functions \( T_{r_a \rightarrow q^*} \), with Markov kernels \( k_a(s, y) = \gamma_a^*(s, y)/\sum_{y' \in \mathcal{Y}} \gamma_a^*(s, y') \), defined for all \( s \in \mathcal{R}_a, y \in \mathcal{Y} \). Essentially, each \( T_{r_a \rightarrow q^*} \) is represented by a \( |\mathcal{R}_a| \times k \) lookup table.
Figure 2: Examples of semi-discrete simplex-vertex optimal transports for $k = 3$. All points in the lower left blue partition are transported to $e_1$, lower right yellow to $e_2$, and upper green to $e_3$. The transports are characterized by a Y-shaped decision boundary.

If the probability mass are not unknown and instead $\text{OPT}(\hat{r}_1, \cdots, \hat{r}_m)$ is solved with the empirical distributions, then the solution provides the optimal transports $T^*_r \rightarrow \hat{q}$ from the empirical $r_a$ to an estimated barycenter $\hat{q}$. We use this transport map to derive a classifier $\hat{h}$ from $f$ (assign arbitrary labels to unseen inputs), and it admits the following guarantees on error rate and fairness:

**Theorem 5** (Finite Sample Generalization, Finite Case). Let $\hat{h}(x, a) = T^*_r \rightarrow \hat{q} \circ f(x, a)$ denote the classifier derived above. Then w.p. at least $1 - \delta$ over the random draws of $S_a$,

$$\sum_{a \in [m]} \text{Err}_a(\hat{h}) - \text{Err}^*_f \leq \sum_{a \in [m]} \sqrt{\frac{2|R_a|^2}{n} \ln \frac{2m|R_a|}{\delta}}$$

and

$$\text{DP Gap}(\hat{h}) \leq \max_{a \in [m]} \sqrt{\frac{R_a^2}{2n} \ln \frac{2m|R_a|}{\delta}}.$$

### 4.2.2 The Continuous Case

When the $r_a$’s are continuous, since the barycenter $\hat{q} \in \mathcal{Q}_k$ (estimated via $\text{OPT}(\hat{r}_1, \cdots, \hat{r}_m)$) is a finite distribution, the simplex-vertex transportation problem from $r_a$ to $\hat{q}$ is semi-discrete, for which it is well-known in optimal transport literature that the optimal transport is deterministic (a.k.a. Monge transport), unique, and specifically, takes the form of a weighted nearest-neighbor function, $s \mapsto e_{\arg \min_i \in [k]}(\|s - e_i\|_1 - \psi_i)$ for some $\psi_i \in \mathbb{R}^k$ (Gangbo and McCann, 1996; Chen et al., 2019). Let $\mathcal{G}_k$ denote the class of such nearest-neighbor functions,

$$\mathcal{G}_k := \left\{ s \mapsto e_{\arg \min_i \in [k]}(\|s - e_i\|_1 - \psi_i) : \psi_i \in \mathbb{R}^k \right\},$$

then given finite samples $S_a$, the optimal transport could be estimated by finding a set of parameters $\{\psi_a\}_{a \in [m]}$ s.t. the associated transport maps $g_a \in \mathcal{G}_k$ minimize the empirical transportation cost from the empirical $\hat{r}_a$’s to $\hat{q}$ (Genevay et al., 2016). Examples of simplex-vertex semi-discrete optimal transports for the $k = 3$ case are illustrated in Fig. 2.

One hurdle, however, is that the existence and uniqueness of the Monge optimal transport for semi-discrete problems are only generally established for strictly convex and superlinear cost functions (Gangbo and McCann, 1996; H1–H3), which our $\ell_1$ cost is not (Ambrosio and Pratelli, 2003). We prove in Appendix D.1 that for our simplex-vertex transportation problem under the $\ell_1$ cost, the Monge optimal transport also exists and is unique.

Therefore, we have properly cast the estimation of the optimal transports $\hat{T}_r \rightarrow \hat{q}$ from the continuous $r_a$’s to the barycenter $\hat{q}$ as empirical risk minimization problems, with $\mathcal{G}_k$ as the set of

---

2Meaning that the measure does not give mass to sets with Hausdorff dimension less than $k - 1$. 11
Algorithm 2: Finite-sample estimate of the barycenter of continuous simplex distributions and the optimal transports to it (corresponds to Lines 4–6 of Algorithm 1)

1 \textbf{function} EstimateBarycenterOT\((S_1, \ldots, S_m)\)
\hspace{1cm} \triangleright \text{each } S_a \text{ contains samples drawn from } r_a := f^\#(\mu_a^X \times \{a\}) , \text{ assumed to be continuous}

2 \hspace{1cm} \textbf{for} a ∈ [m] \textbf{do}
3 \hspace{1.5cm} \hat{r}_a ← \frac{1}{|S_a|} \sum_{s \in S_a} \delta_s \hspace{1cm} \triangleright \hat{r}_a \text{ is empirical distribution of } S_a

4 \hspace{1.5cm} \hat{\gamma}_1^*, \ldots, \hat{\gamma}_m^* ← \text{minimizer of } \text{OPT}(\hat{r}_1, \ldots, \hat{r}_m) \hspace{1cm} \triangleright \text{barycenter of the empirical } \hat{r}_a \text{'s}

5 \hspace{1.5cm} \hat{q} ← \sum_{s \in \text{supp}(\hat{r}_1)} \hat{\gamma}_1^*(s, \cdot) \hspace{1cm} \triangleright \text{each step below is illustrated in Fig. 3}

6 \hspace{1.5cm} \text{Define } v_{ij} := e_j - e_i

7 \hspace{1.5cm} \textbf{for} a ∈ [m] \textbf{do}

8 \hspace{2cm} B_{a,ij} ← \max\{s^\top v_{ij} + 1 : s \in S_a, \hat{\gamma}_a^*(s, e_i) > 0\} \cup \{0\} \hspace{1cm} \triangleright \text{get (offsets of) the boundaries of the set of points transported to each vertex by } \hat{\gamma}_a^*

9 \hspace{2cm} z_a ← \text{arbitrary point in } \bigcap_{i \neq j} \{x \in \mathbb{R}^k : x^\top v_{ij} \geq B_{a,ij} - 1\} \hspace{1cm} \triangleright \text{see Appendix D.1.1 for algorithms; Proposition 14 shows that the intersection is nonempty}

10 \hspace{2cm} \psi_{a,1} ← 0, \text{ and } \psi_{a,i} = 2z^\top v_{i1}, \forall i \neq 1 \hspace{1cm} \triangleright \text{get parameters associated with a mapping in } G_k \text{ whose decision boundaries are centered at } z_a

11 \hspace{2cm} \hat{T}_{r_a \rightarrow \hat{q}} ← (s \mapsto e_{\arg\min_{i \in [k]}(\|s - e_i\|_1 - \psi_{a,i}))} \hspace{1cm} \triangleright \text{estimated simplex-vertex transport}

12 \hspace{1cm} \textbf{return } \hat{q}, \{\hat{T}_{r_a \rightarrow \hat{q}}\}_{a \in [m]}

hypotheses and the empirical transportation costs from the \(\hat{r}_a\)'s as the empirical risks. Curiously, since the solution \(\{\hat{\gamma}_a^*\}_{a \in [m]}\) to \(\text{OPT}(\hat{r}_1, \ldots, \hat{r}_m)\) already gives the discrete transports achieving the minimum empirical risks (Section 4.2.1), so could we leverage the \(\text{OPT}\) solution to get a set of good transport maps directly without needing to solve a new set of optimization problems? Yes: in Algorithm 2 we summarize the procedure for obtaining finite-sample estimates of the barycenter and the optimal transports, and describe between Lines 7–11 the extraction of transports \(\hat{T}_{r_a \rightarrow \hat{q}} \in G_k\) from the OPT solution (illustrated in Fig. 3); specifically, each \(\hat{T}_{r_a \rightarrow \hat{q}}\) from our procedure agrees with the discrete transport \(\hat{\gamma}_a^*\) on all but at most \(k(k - 1)/2\) points in \(S_a\). The derivations are provided in Appendix D.1.1.

Using the transports \(\hat{T}_{r_a \rightarrow \hat{q}} \in G_k\) estimated with \(\text{EstimateBarycenterOT}\) of Algorithm 2, we can derive a deterministic classifier \(\hat{h}\) from \(f\). Thanks to the low complexity of \(G_k\) (Theorem 19), this classifier admits the following finite sample guarantees on error rate and fairness:

**Theorem 6** (Finite Sample Generalization, Continuous Case). Let \(\hat{h}(x, a) = \hat{T}_{r_a \rightarrow \hat{q}} \circ f(x, a)\) denote the classifier derived above. Then w.p. at least \(1 - \delta\) over the random draws of \(S_a\),

\[
\sum_{a \in [m]} \text{Err}_a(\hat{h}) - \text{Err}_a^* \leq O \left( m \left( \sqrt{\frac{k \ln k + \ln 1/\delta}{n}} + \frac{k^2}{n} \right) \right) \quad \text{and}
\]

\[
\text{DP Gap}(\hat{h}) \leq O \left( \sqrt{\frac{k^2 \ln 1/\delta}{n}} + \frac{k^2}{n} \right).
\]

\(^{1}k(k-1)/2 = \binom{k}{2}\) corresponds to the (maximum) number of decision boundaries of each \(g \in G_k\). If the solution \(\{\hat{\gamma}_a^*\}_{a \in [m]}\) to \(\text{OPT}\) is an extremal solution, then combined with a precise tie-breaking rule for the extracted \(\hat{T}_{r_a \rightarrow \hat{q}} \in G_k\), the number of disagreements could be reduced from \(O(k^2)\) to \(k\) (Peyré and Cuturi 2019).
Figure 3: Extract a transport \( g \in \mathcal{G}_3 \) that agrees with the discrete optimal transport (except for points that potentially lie on its boundaries). For illustrative purposes this discrete transport does not split mass.

The first terms of the two expressions above are standard sample complexity for agnostic learning, and the second terms are from the disagreements between the discrete transports \( \hat{\gamma}_a^* \) and the transports \( \hat{T}_{r_a \rightarrow q} \) extracted from them with Algorithm 2 on the dataset \( S_a \) (on average, almost surely).

### 4.2.3 The General Case

For completeness, we briefly discuss the general case where the \( r_a \)'s are neither finite nor purely continuous. One strategy in this scenario is to first construct a (randomized) mapping \( u \) that turns \( r_a \) into a continuous distribution via \( \tilde{r}_a := u \ast r_a \), then find an optimal transport \( T_{\tilde{r}_a \rightarrow q}^* \) from \( \tilde{r}_a \) to the barycenter \( q^* \) (which can be obtained with the procedure in Section 4.2.2 as long as \( \text{supp}(\tilde{r}_a^*) \subseteq \Delta_k \)), and return the mapping given by \( T_{\tilde{r}_a \rightarrow q}^* \circ u \) as a transport from \( r_a \) to \( q^* \) (via \( \tilde{r}_a \)). A careful and precise construction of \( u \) could let the returned mapping recover the optimal transport (Nutz and Wang, 2022), but for simplicity, a lossy approach is described below based on smoothing.

Let \( \rho \) be a continuous distribution with finite first moment, and set \( u \) to the randomized function with Markov kernel \( \mathcal{K}(s, B) = (\rho \ast \delta_s)(B) \) so that the output of \( u \) on any input \( s \) is distributed according to \( s + N, N \sim \rho \). Let \( T_{\tilde{r}_a \rightarrow q}^*, \rho \) denote the optimal transport from \( \tilde{r}_a := u \ast r_a \) to the true barycenter \( q^* \), then we derive a fair classifier \( \tilde{h}_\rho \) from \( f \) using \( T_{\tilde{r}_a \rightarrow q}^*, \rho \circ u \). Note that the smoothing of the \( r_a \) by \( \rho \) introduces noise to the classifier (hence lossy), but the suboptimality due to this noise can be controlled via the “bandwidth” of \( \rho \):

**Proposition 7.** For the fair classifier \( \tilde{h}_\rho \) derived above, \( 0 \leq \sum_{a \in [m]} \text{Err}_a(\tilde{h}_\rho) - \text{Err}_f^* \leq m \cdot \mathbb{E}_{N \sim \rho}[\|N\|_1] \).

For instance, if \( \rho = \text{Laplace}(0, \epsilon I_k) \), then \( \mathbb{E}[\|N\|_1] = k\epsilon \), and the suboptimality introduced by smoothing is no more than \( mk\epsilon \).

### 5 Experiments

We have introduced Algorithm 1 for finding optimal fair classifiers and for post-processing pre-trained predictors for satisfying DP fairness, and provided procedures in Section 4 for obtaining finite-sample estimates. In this section, we verify our proposed method and procedures (Algorithms 1 and 2) on three real-world benchmark datasets, and demonstrate their effectiveness.
For our experiments, on each dataset (described below), we first train a model \( \hat{f} \) to estimate the Bayes regressor via minimizing the squared loss w.r.t. the one-hot labels as in Eq. (6)\(^4\) then learn a post-processing map via estimating the Wasserstein-barycenter \( \hat{q} \) and computing the optimal transports \( \hat{T}_{x \to \hat{q}} \) to the barycenter \( \hat{q} \) from the output distributions of \( \hat{f} \) conditioned on each group. The final post-processed classifier \( (x, a) \mapsto \hat{T}_{x \to \hat{q}} \circ \hat{f}(x, a) \) is evaluated for balanced error rate and the DP gap.

During the training of the regressor \( \hat{f} \), its outputs are not constrained to the simplex for simplicity, but we do project them to the simplex during post-processing and inference because it is a requirement for Algorithm 1. For \textsc{EstimateBarycenterOT} of Algorithm 2, we perform two sets of experiments: one where we treat the output distributions of \( \hat{f} \) as continuous regardless, and another where we apply the smoothing procedure in Section 4.2.3 using \( \text{Laplace}(0, 0.2/|Y| \cdot I) \); results for the latter are relegated to Table 3 in the appendix. Descriptions of the datasets and models are provided below; further experiment details and dataset statistics are included in Appendix E.

Adult \cite{Kohavi1996}. This binary classification task decides whether the annual income of an individual is over or below $50k/year given attributes including gender, race, age, education level, and more, with data collected from the 1994 US Census. We perform three experiments with different sets of sensitive attributes: gender (\(|A| = 2\)), race (5), and their combination (10).

The dataset is pre-processed by normalizing numerical attributes and converting categorical attributes into one-hot representations. The predictor \( \hat{f} \) is a linear regression model. The training set (of 30,162 examples) is split 50%/50% for learning the predictor and the post-processor, respectively.

Communities \cite{RedmondBaveja2002}. The Communities and Crime dataset contains the socioeconomic and crime data of communities in 46 states, and the task is to predict the number of violent crimes per 100k population given attributes ranging from the racial composition of the community, their income and background, and law enforcement resource; the data come from the 1990 US Census and LEMAS survey, and 1995 FBI Uniform Crime Reporting program. The rate of violent crime is binned into five classes, and we treat race as the sensitive attribute by the presence of minorities: a community does not have a significant presence of minorities if White makes up more than 95% of the population, otherwise the largest minority group is considered to have a significant presence (Asian, Black or Hispanic).

The pre-processing and the model are the same as in Adult. Because of the small dataset size of 1,994, results are averaged from 10-fold cross validation (folds are provided by the dataset).

BiosBias \cite{DeArteagaetal2019}. The task is to determine the occupation (out of 28) of female and male individuals by their biography in text; the biographies, occupations and genders are mined from the Common Crawl corpus. Gender is the sensitive attribute, and to which certain occupations such as psychologist and model are correlated in the dataset.

The predictor is a BERT Base (uncased) language model with 110 million parameters fine-tuned from the published checkpoint for three epochs \cite{Devlinetal2019}, whose hyperparameters are included in Appendix E\(^4\). We use the version of BioBias compiled and hosted by \textsc{Ravfogel et al.} \cite{Ravfogeletal2020} containing 255,710 train and 98,344 test examples, and split the training set 95%/5% for learning the predictor and the post-processor, respectively.

\(^4\)Our post-processing procedure is also applicable to models trained with other losses e.g., logistic loss, but performance may be suboptimal without model calibration as discussed in Section 3.2.
Table 2: Results of pre-trained predictors (pre.) and post-processed classifiers (post.) evaluated on test set. Group-balanced accuracy is computed from the balanced error rate in Eq. (1), and DP gap is defined in Eq. (3). Post-processing is performed assuming predictor outputs are continuous.

| Dataset   | Sensitive Attribute | \(|A|\) | \(|Y|\) | Group-Balanced Acc. | DP Gap |
|-----------|---------------------|-------|-------|---------------------|--------|
|           |                     |       |       | Pre. | Post. | ∆ | Pre. | Post. | ∆     |
| Adult     | Gender              | 2     | 2     | 86.35 | 84.34 | -2.33% | 0.1556 | 0.00009 | -99.94% |
|           | Race                | 5     | 2     | 86.97 | 85.55 | -1.63% | 0.1664 | 0.0236 | -85.81% |
|           | G. & R.             | 10    | 2     | 88.22 | 86.45 | -2.00% | 0.2751 | 0.0593 | -78.43% |
| BiosBias  | Gender              | 2     | 28    | 85.86 | 80.68 | -6.03% | 0.2861 | 0.1358 | -52.54% |
| Communities | Race          | 4     | 5     | 67.36 | 60.72 | -9.86% | 0.5321 | 0.1406 | -73.58% |

This experiment is of particular interest because of the increasing popularity of large language models and concerns of their fairness, especially given that the corpus they are pre-trained on are unannotated and may contain historical social bias. Empirical investigations have shown that language models pre-trained on unannotated text from the internet can propagate bias in downstream applications (Bolukbasi et al., 2016; Zhao et al., 2018; Abid et al., 2021).

Results and Analysis. The results on the test set are presented in Table 2 (assuming the output distributions are continuous). Across all datasets and settings—from binary group and binary classification to multigroup and multiclass, and from tabular data to text—our method is effective at reducing the DP gap of the first-stage predictor towards fairness, while incurring only small drops in accuracy. The DP gap of the post-processed classifier is always minimized to ≈ 0 on the train set, and the gap that is present on the test set is due to generalization error (Communities is a small scale dataset, and BiosBias has a considerable number of classes) and/or the violation of the continuous assumption. With the smoothing procedure in Section 4.2.3, the DP gap is further minimized as shown in Table 3. While not explored in our experiments, lower error rates may be possible if the pre-trained predictors were calibrated, as discussed in Section 3.2.

6 Further Related Work

Fairness in Machine Learning. Decision-making systems based on machine learning are influenced by the training data, and could propagate biases present in datasets against historically underrepresented or disadvantaged demographic groups (Barocas and Selbst, 2016; Berk et al., 2021). Towards building fair machine learning systems, various fairness criteria capturing different notions of fairness have been proposed, and are generally categorized into individual (Dwork et al., 2012; Sharifi-Malvajerdi et al., 2019), subgroup (Kearns et al., 2018), and group fairness (Hardt et al., 2016; Kleinberg et al., 2017). While this paper focuses on the group fairness criterion of demographic parity, readers are referred to (Barocas et al., 2019) for discussions and comparisons on the criteria as well as a comprehensive overview of fairness in machine learning.

Bias Mitigation. The concrete mathematical formulations of the fairness criteria enable the design of algorithms for producing fairer systems by satisfying the criteria. The literature contains a wide range of mitigation techniques (Caton and Haas, 2020), including data pre-processing (Calmon et al., 2017; Gordaliza et al., 2019), regularization of training objective (Kamishima et al., 2012;
Learning fair representations (Zemel et al., 2013; Madras et al., 2018; Zhao et al., 2020), and post-processing (Hardt et al., 2016; Pleiss et al., 2017). Our proposed method belongs to this last category.

Bias mitigation typically leads to decreased predictive power of the model due to the inherent trade-offs of model performance and fairness (Calders et al., 2009; Corbett-Davies et al., 2017). As reviewed in Section 1.1, this trade-off, or cost of fairness, is explored in a number of work under different fairness criteria and problem settings.

7 Conclusion

In this paper, we settle the problem of characterizing the minimum cost of fairness for general classification problems under demographic parity. Our theoretical analysis naturally leads to a three-step method for learning optimal fair classifiers. Our method is also applicable to post-processing any pre-trained predictors for satisfying the DP fairness constraint, with finite-sample guarantees on both the accuracy and the DP gap. Technically, we establish the existence and uniqueness of the Monge optimal transport for semi-discrete simplex-vertex transportation problems under the $\ell_1$ cost and the complexity of its function class, which may be of independent interest to the broader community.

Our results add to the line of work that studies the trade-offs between fairness and accuracy, which we believe will not only benefit ML practitioners in the design of fair ML systems but also provide a better understanding of the implications of fairness in machine learning.

Acknowledgements

HZ would like to thank the support from a Facebook research award.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent Anti-Muslim Bias in Large Language Models. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 298–306. 2021.

Alekh Agarwal, Miroslav Dudik, and Zhiwei Steven Wu. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In Proceedings of the 36th International Conference on Machine Learning, pages 120–129, 2019.

Luigi Ambrosio and Aldo Pratelli. Existence and stability results in the $L^1$ theory of optimal transportation. In Optimal Transportation and Applications: Lectures given at the C.I.M.E. Summer School, Held in Martina Franca, Italy, September 2-8, 2001, pages 123–160. 2003.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable Fair Clustering. In Proceedings of the 36th International Conference on Machine Learning, pages 405–413, 2019.

Solon Barocas and Andrew D. Selbst. Big Data’s Disparate Impact. California Law Review, 104 (3):671–732, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. fairmlbook.org, 2019.
Ulrich Bauer, Michael Kerber, Fabian Roll, and Alexander Rolle. A Unified View on the Functorial Nerve Theorem and its Variations, 2022. arXiv:2203.03571 [math].

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research, 50(1):3–44, 2021.

Mathieu Blondel, Akinori Fujino, and Naonori Ueda. Large-Scale Multiclass Support Vector Machine Training via Euclidean Projection onto the Simplex. In 2014 22nd International Conference on Pattern Recognition, pages 1289–1294, 2014.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In Advances in Neural Information Processing Systems, 2016.

Joy Buolamwini and Timnit Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pages 77–91, 2018.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building Classifiers with Independency Constraints. In 2009 IEEE International Conference on Data Mining Workshops, pages 13–18, 2009.

Flavio P. Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized Pre-Processing for Discrimination Prevention. In Advances in Neural Information Processing Systems, 2017.

Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey, 2020. arXiv:2010.04053 [cs, stat].

Yucheng Chen, Matus Telgarsky, Chao Zhang, Bolton Bailey, Daniel Hsu, and Jian Peng. A Gradual, Semi-Discrete Approach to Generative Network Training via Explicit Wasserstein Minimization. In Proceedings of the 36th International Conference on Machine Learning, pages 1071–1080, 2019.

Jianfeng Chi, Jian Shen, Xinyi Dai, Weinan Zhang, Yuan Tian, and Han Zhao. Towards Return Parity in Markov Decision Processes. In Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, pages 1161–1178, 2022.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair Clustering Through Fairlets. In Advances in Neural Information Processing Systems, 2017.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair Regression with Wasserstein Barycenters. In Advances in Neural Information Processing Systems, pages 7321–7331, 2020.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 797–806, 2017.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in Bios:
A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantee in multi-class classification, 2022. *arXiv:2109.13642 [math, stat]*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268, 2015.

Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

Solenne Gaucher, Nicolas Schreuder, and Evgenii Chzhen. Fair learning with Wasserstein barycenters for non-decomposable performance measures, 2022. *arXiv:2209.00427 [cs, math, stat]*.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic Optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems*, 2016.

Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online Learning with an Unknown Fairness Metric. In *Advances in Neural Information Processing Systems*, 2018.

Paula Gordaliza, Eustasio del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining Fairness using Optimal Transport Theory. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2357–2365, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330, 2017.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, 2016.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein Fair Classification. In *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, pages 862–872, 2020.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems*, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Fair Algorithms for Infinite and Contextual Bandits, 2017. *arXiv:1610.09559 [cs]*.
Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2564–2572, 2018.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference*, pages 43:1–43:23, 2017.

Ron Kohavi. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

Thibaut Le Gouic, Jean-Michel Loubes, and Philippe Rigollet. Projection to Fairness in Statistical Learning, 2020. arXiv:2005.11720 [cs, math, stat].

Peizhao Li, Han Zhao, and Hongfu Liu. Deep Fair Clustering for Visual Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9067–9076, 2020.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393, 2018.

B. K. Natarajan. On Learning Sets and Functions. *Machine Learning*, 4(1):67–97, 1989.

Marcel Nutz and Ruodu Wang. The Directional Optimal Transport. *The Annals of Applied Probability*, 32(2):1400–1420, 2022.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport: With Applications to Data Science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, 2017.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, 2020.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Saeed Sharifi-Malvajerdi, Michael Kearns, and Aaron Roth. Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems*, 2019.
Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning Controllable Fair Representations. In Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, pages 2164–2173, 2019.

Edwin H. Spanier. Algebraic Topology. Springer, 1981.

Sahil Verma and Julia Rubin. Fairness Definitions Explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), 2018.

Cédric Villani. Topics in Optimal Transportation. American Mathematical Society, 2003.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Cauwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pages 962–970, 2017.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. FA*IR: A Fair Top-k Ranking Algorithm. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1569–1578, 2017.

Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In Proceedings of the 30th International Conference on Machine Learning, pages 325–333, 2013.

Han Zhao and Geoffrey J. Gordon. Inherent Tradeoffs in Learning Fair Representations. Journal of Machine Learning Research, 23(57):1–26, 2022.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional Learning of Fair Representations. In International Conference on Learning Representations, 2020.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 15–20, 2018.
A Additional Discussions

A.1 Weighted Wasserstein-Barycenter

To extend our results to arbitrary group-weighted classification error of

\[
\sum_{a \in [m]} w_a \cdot \text{Err}_a(h) := \sum_{a \in [m]} w_a \cdot \mathbb{P}(h(X, a) \neq Y \mid A = a)
\]

with \(w \in \mathbb{R}_{\geq 0}^m\), it suffices to establish a version of Theorem 1 for this weighted error.

**Theorem 8.** Let \(\mu\) be given, let \(f^* : \mathcal{X} \times A \rightarrow \Delta_k\) denote the Bayes regressor on \(\mu\), \(r^*_a := f^*_{\mu_a}^*(\mu_a \times \{a\})\), and \(Q_k\) the set of distributions on \(\mathcal{Y} = \{e_1, \cdots, e_k\}\). Fix \(w \in \mathbb{R}_{\geq 0}^m\). Then

\[
\min_{h: \text{DP holds}} \sum_{a \in [m]} w_a \cdot \text{Err}_a(h) = \frac{1}{2} \min_{q \in Q_k} \sum_{a \in [m]} w_a \cdot W_1(r^*_a, q).
\]

**Proof.** Lemma 2 implies that for any fixed \(q \in Q_k\), for all \(a \in [m]\),

\[
\min_{h: \text{DP holds}} w_a \cdot \text{Err}_a(h) = w_a \min_{h: \text{DP holds}} \mathbb{P}(h(X, a) \neq Y \mid A = a) = \frac{1}{2} w_a \cdot W_1(r^*_a, q).
\]

By optimizing the components \(h(\cdot, a)\) independently,

\[
\sum_{a \in [m]} \min_{h: \text{DP holds}} w_a \cdot \mathbb{P}(h(X, a) \neq Y \mid A = a) = \min_{q \in Q_k} \sum_{a \in [m]} w_a \cdot \mathbb{P}(h(X, a) \neq Y \mid A = a),
\]

so

\[
\min_{h: \text{DP holds}} \sum_{a \in [m]} w_a \cdot \text{Err}_a(h) = \min_{q \in Q_k} \sum_{a \in [m]} w_a \cdot \text{Err}_a(h)
\]

\[
= \frac{1}{2} \min_{q \in Q_k} \sum_{a \in [m]} w_a \cdot W_1(r^*_a, q).
\]

\(\blacksquare\)

A.2 Reduction to TV-Barycenter

**Proposition 9.** Let realizable \(\mu\) be given, i.e., \(\min_h \text{Err}_a(h) = 0, \forall a \in [m]\), let \(f^* : \mathcal{X} \times A \rightarrow \Delta_k\) denote the Bayes regressor on \(\mu\), \(r^*_a := f^*_{\mu_a}^*(\mu_a \times \{a\})\), and \(Q_k\) the set of distributions on \(\mathcal{Y} = \{e_1, \cdots, e_k\}\). Then

\[
\frac{1}{2} \min_{q \in Q_k} \sum_{a \in [m]} W_1(r^*_a, q) = \min_{q \in Q_k} \sum_{a \in [m]} d_{\text{TV}}(p^*_a, q),
\]

where \(p^*_a\) is the true class distribution conditioned on group \(a\), \(p^*_a(e_i) := \mathbb{P}_\mu(Y = e_i \mid A = a)\).

**Proof.** First, realizable means that there exists a deterministic labeling function \(f : \mathcal{X} \times A \rightarrow \mathcal{Y}\) s.t. \(Y = f(X, A)\), in which case \(f^* = f\), so we have \(r^*_a(e_i) = \mathbb{P}(f(X, a) = e_i \mid A = a) = p^*_a(e_i)\). Because now \(\text{supp}(r^*_a) \subseteq \mathcal{Y}\), the \(\ell_1\) cost between any \(s \in \text{supp}(r^*_a)\) and \(y \in \mathcal{Y}\) reduces to \(\|s - y\|_1 = 2 \cdot 1(s \neq y)\).
To establish the result, it suffices to show that \( \forall q \in Q_k, \exists q' \in Q_k \) s.t. the following equality holds:

\[
\frac{1}{2} \sum_{a \in [m]} W_1(r^*_a, q) = \sum_{a \in [m]} d_{TV}(p^*_a, q'),
\]

and conversely, \( \forall q' \in Q_k, \exists q \in Q_k \) s.t. Eq. (8) holds.

Observe that

\[
\frac{1}{2} W_1(r^*_a, q) = \frac{1}{2} \min_{\gamma \in \Gamma(r^*_a, q)} \int_{\mathcal{Y} \times \mathcal{Y}} ||s - y||_1 \, d\gamma(s, y)
\]

\[
= \min_{\gamma \in \Gamma(r^*_a, q)} \sum_{s \neq y} \gamma(s, y) = 1 - \max_{\gamma \in \Gamma(r^*_a, q)} \sum_{i \in [k]} \gamma(e_i, e_i).
\]

Let \( q \in Q_k \) be arbitrary, we show that the choice of \( q' = q \) satisfies Eq. (8). Note that

\[
\forall \gamma \in \Gamma(r^*_a, q), \gamma(e_i, e_i) \leq \min(q'(e_i), p^*_a(e_i)) \quad \text{and} \quad \exists \gamma \in \Gamma(r^*_a, q), \gamma(e_i, e_i) \geq \min(q'(e_i), p^*_a(e_i)).
\]

and these statements imply that the maximizing choice is \( \gamma(e_i, e_i) = \min(q'(e_i), p^*_a(e_i)) \), so

\[
\frac{1}{2} W_1(r^*_a, q) = 1 - \sum_{i \in [k]} \min(q'(e_i), p^*_a(e_i)) = \sum_{i \in [k]} \left( q'(e_i) - \min(q'(e_i), p^*_a(e_i)) \right)
\]

\[
= \sum_{i \in [k]} \max(0, q'(e_i) - p^*_a(e_i)) = \frac{1}{2} \sum_{i \in [k]} |p^*_a(e_i) - q'(e_i)| = d_{TV}(p^*_a, q'),
\]

where the last equality is due to \( \sum_{i \in [k]} (q'(e_i) - p^*_a(e_i)) = 0 \). Conversely, let \( q' \in Q_k \) be arbitrary, then the choice of \( q = q' \) satisfies Eq. (8) by the same reasoning.

\section*{B Omitted Proofs from Section 3}

The proofs make use of the following result that rewrites the error rate as an integral over a coupling:

\begin{lemma}
Let \( \mu \) with only one demographic group be given, let \( f^*(x) := \mathbb{E}_\mu[Y \mid X = x] \) denote the Bayes regressor on \( \mu \). Then the error rate of any randomized classifier \( h : \mathcal{X} \to \mathcal{Y} \) is given by

\[
\text{Err}(h) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} ||s - y||_1 \, d\mathbb{P}(f^*(X) = s, h(X) = y) = \frac{1}{2} \mathbb{E}[||h(X) - f^*(X)||_1].
\]

Note that the joint distribution of \((f^*(X), h(X))\) is a coupling belonging to \( \Gamma(f^*_\mu X, h^*_\mu X) \).
\end{lemma}

\begin{proof}
1 - \text{Err}(h) = 1 - \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h(X) = Y)
\]

\[
= \int_{\Delta_k} \sum_{i \in [k]} \mathbb{P}(h(X) = e_i, Y = e_i, f^*(X) = s) \, ds
\]

\[
= \int_{\Delta_k} \sum_{i \in [k]} \mathbb{P}(h(X) = e_i, Y = e_i \mid f^*(X) = s) \cdot \mathbb{P}_\mu(f^*(X) = s) \, ds
\]

\[
= \int_{\Delta_k} \sum_{i \in [k]} \mathbb{P}(h(X) = e_i \mid f^*(X) = s) \cdot \mathbb{P}_\mu(Y = e_i \mid f^*(X) = s) \cdot \mathbb{P}_\mu(f^*(X) = s) \, ds
\]

\[
= \int_{\Delta_k} \sum_{i \in [k]} s_i \cdot \mathbb{P}(h(X) = e_i, f^*(X) = s) \, ds,
\]

\end{proof}
where line 4 follows from $X \perp Y$ given $f^*(X)$, since $f^*(X) = \mathbb{E}_\mu[Y \mid X]$ fully specifies the pmf of $Y$ conditioned on $X$. So

$$\begin{align*}
\text{Err}(h) &= \int_{\Delta_k} \sum_{i \in [k]} (1 - s_i) \cdot \mathbb{P}(h(X) = e_i, f^*(X) = s) ds \\
&= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - e_i\|_1 \cdot \mathbb{P}(h(X) = e_i, f^*(X) = s) ds \\
&= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 d\mathbb{P}(f^*(X) = s, h(X) = y) \\
&= \frac{1}{2} \mathbb{E}[\|h(X) - f^*(X)\|_1],
\end{align*}$$

where the second equality is due to an identity stated in Eq. (10) in a later section.

\[\blacksquare\]

**Lemma 11 (Extended Version of Lemma 2).** Let $\mu$ with only one demographic group be given, let $f^*(x) := \mathbb{E}_\mu[Y \mid X = x]$ denote the Bayes regressor on $\mu$, and define $r^* := f^*\#\mu^X$. Fix $q \in \mathcal{Q}_k$.

Then for any randomized classifier $h : X \rightarrow \mathcal{Y}$ with Markov kernel $K$ satisfying $h\#\mu^X \overset{d}{=} q$, the coupling $\gamma \in \Gamma(r^*, q)$ given by $\gamma(s, y) = \int_{f^*^{-1}(s)} K(x, y) d\mu^X(x)$ satisfies

$$\text{Err}(h) := \mathbb{P}(h(X) \neq Y) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 d\gamma(s, y). \quad (9)$$

Conversely, for any $\gamma \in \Gamma(r^*, q)$, the randomized classifier $h$ with Markov kernel $K(x, B) = \gamma(f^*(x), B)/\gamma(f^*(x), \mathcal{Y})$ satisfies $h\#\mu^X \overset{d}{=} q$ and Eq. (9).

**Proof.** Let a randomized classifier $h$ satisfying $h\#\mu^X \overset{d}{=} q$ and associated with Markov kernel $K$ be given. We first verify that the constructed coupling belongs to $\Gamma(r^*, q)$:

$$\begin{align*}
\int_{\mathcal{Y}} \gamma(s, y) dy &= \int_{\mathcal{Y}} \int_{f^*^{-1}(s)} K(x, y) d\mu^X(x) dy = \int_{f^*^{-1}(s)} \int_{\mathcal{Y}} K(x, y) dy d\mu^X(x) \\
&= \int_{f^*^{-1}(s)} d\mu^X(x) = \mathbb{P}_\mu(f^*(X) = s) = r^*(s),
\end{align*}$$

and

$$\begin{align*}
\int_{\Delta_k} \gamma(s, y) ds &= \int_{\Delta_k} \int_{f^*^{-1}(s)} K(x, y) d\mu^X(x) ds = \int_{\mathcal{X}} K(x, y) d\mu^X(x) \\
&= \int_{\mathcal{X}} \mathbb{P}(h(X) = y \mid X = x) d\mu^X(x) = \mathbb{P}(h(X) = y) = q(y).
\end{align*}$$

23
Next, by Lemma 10,

$$\text{Err}(h) = \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \cdot P(f^*(X) = s, h(X) = y) \, ds$$

$$= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \int_x P(f^*(X) = s, h(X) = y, X = x) \, dx \, ds$$

$$= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \int_{f^*-1(s)} P(h(X) = y, X = x) \, dx \, ds$$

$$= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \int_{f^*-1(s)} K(x, y) \cdot P(\mu(X = x)) \, dx \, ds$$

$$= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \int_{f^*-1(s)} K(x, y) \cdot \mu(X = x) \, dx \, ds$$

$$= \frac{1}{2} \int_{\Delta_k} \sum_{i \in [k]} \|s - y\|_1 \gamma(s, y) \, ds = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 d\gamma(s, y),$$

where line 3 is due to $P(f^*(X) = s, h(X) = y, X = x) = \mathbb{1}(f^*(x) = s) \cdot P(h(X) = y, X = x)$ for all $s \in \Delta_k$, $y \in \mathcal{Y}$, $x \in \mathcal{X}$.

On the other hand, let a coupling $\gamma \in \Gamma(r^*, q)$ be given. We show that the Markov kernel of the constructed randomized classifier $\tilde{h}$ satisfies $\gamma(s, y) = \int_{f^*-1(s)} K(x, y) \, d\mu_X(x)$, whereby Eq. (9) follows from the same arguments used in the previous part: let $s \in \Delta_k$ and $y \in \mathcal{Y}$ be given, let $x' \in f^*-1(s)$ be arbitrary, then

$$\gamma(s, y) = K(x', y) \cdot \gamma(s, \mathcal{Y}) = K(x', y) \cdot r^*(s)$$

$$= K(x', y) \int_{x \in f^*-1(s)} d\mu_X(x) = \int_{x \in f^*-1(s)} K(x, y) \, d\mu_X(x),$$

where the last line follows because $K(x, y)$ is constant for all $x \in f^*-1(s)$ by construction.

**Proof of Theorem 3.** Note that the Markov kernel of the randomized classifier $\tilde{h}$ is given by $K(x, B) = \gamma_a^*(f^*(x), B) / \gamma_a^*(f^*(x), \mathcal{Y})$, where $\gamma_a^* \in \Gamma(r_a^*, q)$ is an optimal transport between $r_a^*$ and $q$. So $\gamma_a^*(s, \mathcal{Y}) = r_a^*(s) := \int_{\mathcal{Y}} d\mu_X(x, a = s)$ and $\gamma_a^*(\Delta_k, y) = q(y)$ for all $s \in \Delta_k$ and $y \in \mathcal{Y}$.

Firstly, define $f^*-1(s) = \{x \in \mathcal{X} : f^*(x, a) = s\}$, then Lemma 11 implies that $\text{Err}_a(\tilde{h}) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 d\gamma_a(s, y)$ with

$$\gamma_a(s, y) := \int_{f^*-1(s)} K(x, y) \, d\mu_X(x) = \int_{f^*-1(s)} \frac{\gamma_a^*(f^*(x, a), y)}{\gamma_a^*(f^*(x, a), \mathcal{Y})} \, d\mu_X(x)$$

$$= \frac{\gamma_a^*(s, y)}{\gamma_a^*(s, \mathcal{Y})} \int_{f^*-1(s)} d\mu_X(x) = \gamma_a^*(s, y),$$

where the second equality is due to $\gamma_a^*(f^*(x, a), y) = \gamma_a^*(s, y)$ constant for all $x \in f^*-1(s)$, so $\text{Err}_a(\tilde{h}) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 d\gamma_a(s, y) = \frac{1}{2} W_1(r_a^*, q)$ because $\gamma_a^*$ is an optimal transport.
Secondly, for all \( y \in Y \),
\[
\mathbb{P}(\bar{h}(X, a) = y \mid A = a) = \int_X \mathbb{P}(\bar{h}(X, a) = y \mid X = x, A = a) \cdot \mathbb{P}_\mu(X = x \mid A = a) \, dx
\]
\[
= \int_X \gamma_a^*(f^*(x, a), y) \, d\mu_a^X(x) = \int_{\Delta_k} \gamma_a^*(s, y) \int_{f_a^{-1}(s)} d\mu_a^X(x) \, ds
\]
\[
= \int_{\Delta_k} \gamma_a^*(s, y) \, ds = \gamma_a^*(\Delta_k, y) = q(y).
\]

\[\square\]

Proof of Theorem 4. The randomized classifier returned by \textsc{PostProcess}($\tilde{f}; \mu^X$) of Algorithm 1 is $\bar{h}(x, a) = T_{\tilde{r}_a \rightarrow \tilde{q}^*} \circ \tilde{f}(x, a)$, where $T_{\tilde{r}_a \rightarrow \tilde{q}^*}$ is the optimal transport from $\tilde{r}_a := \tilde{f}_a^*(\mu^X \times \{a\})$ to $\tilde{q}^*$, and $\tilde{q}^*$ is the barycenter of the $\tilde{r}_a$’s. Define $\epsilon_a := \mathbb{E}_{\mu^X} [||\tilde{f}(X, a) - f^*(X, a)||_1]$.

By Lemma 10 and triangle inequality,
\[
\sum_{a \in [m]} \text{Err}_a(\bar{h}) = \frac{1}{2} \sum_{a \in [m]} \mathbb{E} \left[ ||T_{\tilde{r}_a \rightarrow \tilde{q}^*} \circ \tilde{f}(X, a) - f^*(X, a)||_1 \mid A = a \right]
\]
\[
\leq \frac{1}{2} \sum_{a \in [m]} \mathbb{E} \left[ ||T_{\tilde{r}_a \rightarrow \tilde{q}^*} \circ \tilde{f}(X, a) - \tilde{f}(X, a)||_1 + ||\tilde{f}(X, a) - f^*(X, a)||_1 \mid A = a \right]
\]
\[
\leq \frac{1}{2} \sum_{a \in [m]} W_1(\tilde{r}_a, \tilde{q}^*) + \frac{\epsilon_a}{2} = \frac{1}{2} \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} W_1(\tilde{r}_a, q) + \frac{\epsilon_a}{2},
\]
where the inequality on line 3 is due to the optimal transport $T_{\tilde{r}_a \rightarrow \tilde{q}^*}$, and the equality to the barycenter $\tilde{q}^*$. So
\[
\sum_{a \in [m]} \text{Err}_a(\bar{h}) - \text{Err}^* \leq \frac{1}{2} \left( \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} W_1(\tilde{r}_a, q) - \sum_{a \in [m]} W_1(r_a^*, q^*) \right) + \frac{\epsilon_a}{2}
\]
\[
\leq \frac{1}{2} \left( \sum_{a \in [m]} W_1(\tilde{r}_a, q^*) - \sum_{a \in [m]} W_1(r_a^*, q^*) \right) + \frac{\epsilon_a}{2}
\]
\[
\leq \frac{1}{2} \sum_{a \in [m]} W_1(\tilde{r}_a, r_a^*) + \frac{\epsilon_a}{2} \leq \epsilon_a,
\]
where the last inequality is because the joint distribution $\gamma_a$ of $(\tilde{f}(X, a), f^*(X, a))$ conditioned on $A = a$ is a coupling belonging to $\Gamma(\tilde{r}_a, r_a^*)$, and $\int_{\Delta_k \times Y} \delta s - y_1 \, d\gamma_a(s, y) = \int_X \|\tilde{f}(x, a) - f^*(x, a)\|_1 \, d\mu^X_a(x) = \epsilon_a$.

On the other hand, again by Lemma 10
\[
\sum_{a \in [m]} \text{Err}_a(\bar{h}) = \frac{1}{2} \sum_{a \in [m]} \int_{\Delta_k \times Y} \|s - y\|_1 \, d\mathbb{P}(f^*(X, a) = s, T_{\tilde{r}_a \rightarrow \tilde{q}^*} \circ \tilde{f}(X, a) = y \mid A = a)
\]
\[
\geq \frac{1}{2} \sum_{a \in [m]} W_1(r_a^*, \tilde{q}^*) \geq \frac{1}{2} \min_{q \in \mathcal{Q}_k} \sum_{a \in [m]} W_1(r_a^*, q) = \text{Err}^*,
\]
where the first inequality follows from the joint distribution of $(f^*(X, a), T_{\tilde{r}_a \rightarrow \tilde{q}^*} \circ \tilde{f}(X, a))$ conditioned on $A = a$ being a coupling belonging to $\Gamma(r_a^*, \tilde{q}^*)$, and the definition of Wasserstein distance. \[\square\]
C Omitted Proofs from Section 4

For the finite sample generalization bound in the finite case, we use the following \( \ell_1 \) convergence result of empirical distributions:

**Lemma 12.** Let \( p \) be a finite distribution on \( \mathcal{X} \), and \( x_1, \ldots, x_n \sim p \) i.i.d. samples, with which the empirical distribution \( \hat{p}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i} \) is formed. Then w.p. at least \( 1 - \delta \) over the random draw of the \( x_i \)'s, \( \|p - \hat{p}_n\|_1 \leq \sqrt{\frac{|X|^2}{2n} \ln \frac{2|X|}{\delta}}. \)

**Proof.** By Hoeffding’s inequality and union bound, w.p. at least \( 1 - \delta \), \( |\hat{p}_n(x) - p(x)| \leq \sqrt{\frac{1}{2n} \ln \frac{2|X|}{\delta}} \) for all \( x \in \mathcal{X} \), and the result follows from summing over \( x \in \mathcal{X} \).

**Proof of Theorem 5.** Firstly, by Lemma 10

\[
\text{Err}_a(\hat{h}) = \frac{1}{2} \sum_{s \in \mathcal{R}_a} \sum_{y \in \mathcal{Y}} \|s - y\|_1 \cdot r_a(s) \cdot P(T_{\hat{r}_a \rightarrow \hat{q}}^*(s) = y) 
\leq \frac{1}{2} \sum_{s \in \mathcal{R}_a} \sum_{y \in \mathcal{Y}} \|s - y\|_1 \cdot (\hat{r}_a(s) + |r_a(s) - \hat{r}_a(s)|) \cdot P(T_{\hat{r}_a \rightarrow \hat{q}}^*(s) = y) 
\leq \frac{1}{2} W_1(\hat{r}_a, \hat{q}) + \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)|,
\]

where line 3 is due to the optimal transport \( T_{\hat{r}_a \rightarrow \hat{q}}^* \) from \( \hat{r}_a \) to \( \hat{q} \). So

\[
\sum_{a \in [m]} \text{Err}_a(\hat{h}) - \text{Err}_a^* \leq \sum_{a \in [m]} \left( \frac{1}{2} (W_1(\hat{r}_a, \hat{q}) - W_1(r_a, q^*)) + \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)| \right) 
\leq \sum_{a \in [m]} \left( \frac{1}{2} (W_1(\hat{r}_a, q^*) - W_1(r_a, q^*)) + \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)| \right) 
\leq \sum_{a \in [m]} \left( \frac{1}{2} W_1(\hat{r}_a, r_a) + \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)| \right),
\]

where line 2 is due to \( \hat{q} \) being the Wasserstein-barycenter of the \( \hat{r}_a \)'s. The second term is the TV distance between \( r_a \) and \( \hat{r}_a \), which also upper bounds the \( W_1 \) distance between them in the first term: note that \( \exists \gamma \in \Gamma(\hat{r}_a, r_a) \) s.t. \( \gamma(s, s) \geq \min(\hat{r}_a(s), r_a(s)) \) for all \( s \in \mathcal{R}_a \), so by the same arguments in the proof of Proposition 9

\[
\frac{1}{2} W_1(\hat{r}_a, r_a) = \frac{1}{2} \inf_{\gamma \in \Gamma(\hat{r}_a, r_a)} \int \|\hat{s} - s\|_1 d\gamma(\hat{s}, s) \leq \min_{\Gamma(\hat{r}_a, r_a)} \sum_{\hat{s} \neq s} d\gamma(\hat{s}, s) 
= 1 - \max_{\Gamma(\hat{r}_a, r_a)} \sum_{s} \gamma(s, s) = 1 - \sum_{s \in \mathcal{R}_a} \min(\hat{r}_a(s), r_a(s)) 
= \frac{1}{2} \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)|.
\]

The result follows from Lemma 12 and a union bound.
Secondly, note that for all \( a \in [m] \) and \( y \in \mathcal{Y} \),
\[
\sum_{y \in \mathcal{Y}} |\mathbb{P}(\hat{h}(X, a) = y \mid A = a) - \hat{q}(y)| = \sum_{y \in \mathcal{Y}} \sum_{s \in \mathcal{R}_a} r_a(s) \cdot \mathbb{P}(T^*_{\hat{r}_a \rightarrow \hat{q}}(s) = y) - \sum_{s \in \mathcal{R}_a} \hat{r}_a(s) \cdot \mathbb{P}(T^*_{\hat{r}_a \rightarrow \hat{q}}(s) = y) \\
\leq \sum_{y \in \mathcal{Y}} \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)| \cdot \mathbb{P}(T^*_{\hat{r}_a \rightarrow \hat{q}}(s) = y) \\
= \sum_{s \in \mathcal{R}_a} |r_a(s) - \hat{r}_a(s)|,
\]
where the first line follows from the transport \( T^*_{\hat{r}_a \rightarrow \hat{q}} \) from \( \hat{r}_a \) to \( \hat{q} \). Then the result follows from Lemma 12 and union bounds. In particular, we used the fact that for all \( a, a' \in [m] \) and \( y \in \mathcal{Y} \),
\[
|\mathbb{P}(\hat{h}(X, a) = y \mid A = a) - \mathbb{P}(\hat{h}(X, a') = y \mid A = a')| \leq |\mathbb{P}(\hat{h}(X, a) = y \mid A = a) - \hat{q}(y)| + |\mathbb{P}(\hat{h}(X, a) = \hat{q}(y)) - \hat{q}(y)|.
\]

For the finite sample generalization bound in the continuous case, we require the following agnostic PAC learning bound with Natarajan dimension as the complexity measure [Natarajan, 1989]:

**Lemma 13** (Shalev-Shwartz and Ben-David, 2014, Theorem 29.3). Let a function class \( \mathcal{H} \subset \{1, \cdots, k\}^\mathcal{X} \) and a loss function \( \ell : \mathcal{H} \times \mathcal{X} \to \mathbb{R}_{\geq 0} \) be given. Let \( p \) be a distribution on \( \mathcal{X} \), and \( x_1, \cdots, x_n \sim p \) i.i.d. samples. Then w.p. at least \( 1 - \delta \) over the random draw of the \( x_i \)'s, \( \forall h \in \mathcal{H}, \)
\[
\left| \mathbb{E}_{X \sim p}[\ell(h, X)] - \frac{1}{n} \sum_{i=1}^{n} \ell(h, x_i) \right| \leq \sqrt{\frac{C}{n} \left( d_N(\mathcal{H}) \ln k + \ln \frac{1}{\delta} \right)},
\]
where \( C > 0 \) is a universal constant, and \( d_N(\mathcal{H}) \) is the Natarajan dimension of the hypothesis class \( \mathcal{H} \).

**Proof of Theorem 7**. Firstly, by Lemma 10
\[
\text{Err}_a(\hat{h}) = \frac{1}{2} \mathbb{E}_{S \sim r^*_a}[\| \hat{T}_{r_a \rightarrow \hat{q}}(S) - S \|_1].
\]
Here, \( \hat{T}_{r_a \rightarrow \hat{q}} \in \mathcal{G}_k \), for which we show in Theorem 19 that \( d_N(\mathcal{G}_k) = k - 1 \). Then by Lemma 13, w.p. at least \( 1 - \delta \),
\[
\left| \mathbb{E}_{S \sim r^*_a}[\| \hat{T}_{r_a \rightarrow \hat{q}}(S) - S \|_1] - \frac{1}{n} \sum_{j=1}^{n} \| \hat{T}_{r_a \rightarrow \hat{q}}(s_{a,j}) - s_{a,j} \|_1 \right| \leq O \left( \sqrt{\frac{k \ln k + \ln 1/\delta}{n}} \right)
\]
with \( \ell(T, s) := \| T(s) - s \|_1 \).

Now, we want to relate the empirical loss in the second term to the optimal transportation cost of the empirical \( \hat{r}_a \) to \( \hat{q} \), which is achieved by the discrete transport \( T^*_{\hat{r}_a \rightarrow \hat{q}} \) obtained from the OPT solution \( \{ \hat{\gamma}_a \}_{a \in [m]} \). Because our particular \( \hat{T}_{r_a \rightarrow \hat{q}} \) is extracted from \( T^*_{\hat{r}_a \rightarrow \hat{q}} \) using the procedure described in Section 4.2.2 by Corollary 15, they agree on \( S_a \) except for points that lie on the boundaries of \( T^*_{\hat{r}_a \rightarrow \hat{q}} \). The total number of boundaries is \( \binom{k}{2} = k(k - 1)/2 \), but because the set of points on the boundaries have measure zero by the continuity of \( r_a \), almost surely no two points
lie on the same boundary, which means that \( \hat{T}_{r_a \rightarrow \hat{q}} \) agrees with \( T^*_{r_a \hat{q}} \) on all but \( O(k^2) \) points in \( S_a \). Let the set of points on which they disagree be denoted by \( S'_a \), then

\[
\frac{1}{n} \sum_{j=1}^{n} \left| \frac{1}{n} \sum_{(s,y) \in S_a \times Y} \| \hat{T}_{r_a \rightarrow \hat{q}}(s_{a,j}) - s_{a,j} \|_1 - \sum_{(s,y) \in S_a \times Y} \| s - y \|_1 \cdot \hat{\gamma}^*(s, y) \right|
\]

\[
= \sum_{(s,y) \in S_a \times Y} \| s - y \|_1 \cdot ((\text{Id} \times \hat{T}_{r_a \rightarrow \hat{q}})(s, y) - \sum_{(s,y) \in S_a \times Y} \| s - y \|_1 \cdot \hat{\gamma}^*(s, y))
\]

\[
\leq \sum_{i \in [k]} \sum_{s \in S_a} \| s - e_i \|_1 \cdot ((\text{Id} \times \hat{T}_{r_a \rightarrow \hat{q}})(s, e_i) - \sum_{s \in S_a} \| s - e_i \|_1 \cdot \hat{\gamma}^*(s, e_i))
\]

\[
\leq 2 \sum_{i \in [k]} \sum_{s \in S_a} \left( (\text{Id} \times \hat{T}_{r_a \rightarrow \hat{q}})(s, e_i) - \hat{\gamma}^*(s, e_i) \right) = O(k^2/n).
\]

Let \( T^*_{r_a \rightarrow q^*} \in G_k \) denote the Monge optimal transport from \( r_a \) to \( q^* \), which exists due to Lemma \[16\] and define \( q'(e_i) = \frac{1}{n} \sum_{s \in S_a} \mathbb{I}(T^*_{r_a \rightarrow q^*}(s) = e_i) \) for all \( i \in [k] \). Then \( (\text{Id} \times T^*_{r_a \rightarrow q^*})\hat{\gamma}_a \in \Gamma(\hat{r}_a, q') \). Define for simplicity \( E := O(\sqrt{(k \ln k + \ln 1/\delta)/n + k^2/n}) \); it follows from above that

\[
\sum_{a \in [m]} \text{Err}_a(\hat{h}) - \text{Err}^f
\]

\[
\leq mE + \frac{1}{2} \sum_{a \in [m]} \left( W_1(\hat{r}_a, \hat{q}) - W_1(r_a, q^*) \right)
\]

\[
= mE + \frac{1}{2} \sum_{a \in [m]} \left( W_1(\hat{r}_a, \hat{q}) - W_1(\hat{r}_a, q^*) + W_1(\hat{r}_a, q^*) - W_1(r_a, q^*) \right)
\]

\[
\leq mE + \frac{1}{2} \sum_{a \in [m]} \left( W_1(\hat{r}_a, q^*) - W_1(r_a, q^*) \right)
\]

\[
\leq mE + \frac{1}{2} \sum_{a \in [m]} \left( \frac{1}{n} \sum_{j=1}^{n} \left\| T^*_{r_a \rightarrow q^*}(s_{a,j}) - s_{a,j} \right\|_1 - E_{S \sim r_a^\#} \left[ \left\| T^*_{r_a \rightarrow q^*}(S) - S \right\|_1 \right] \right)
\]

where line 3 is due to the barycenter \( \hat{q} \) of the \( \hat{r}_a \)'s, and line 4 to \( T^*_{r_a \rightarrow q^*} \) being a transport from \( \hat{r}_a \) (sample) to \( q' \) and an optimal transport from \( r_a \) (population) to \( q^* \). Applying Lemma \[13\] again, we get w.p. at least \( 1 - \delta \),

\[
\frac{1}{n} \sum_{j=1}^{n} \left\| T^*_{r_a \rightarrow q^*}(s_{a,j}) - s_{a,j} \right\|_1 - E_{S \sim r_a^\#} \left[ \left\| T^*_{r_a \rightarrow q^*}(S) - S \right\|_1 \right] \leq O\left( \sqrt{\frac{k \ln k + \ln 1/\delta}{n}} \right)
\]

hence with a union bound we conclude

\[
\sum_{a \in [m]} \text{Err}_a(\hat{h}) - \text{Err}^f \leq O\left( m \left( \sqrt{\frac{k \ln k + \ln 1/\delta}{n}} + \frac{k^2}{n} \right) \right).
\]

Secondly and similarly, for all \( a \in [m] \) and \( i \in [k] \), w.p. at least \( 1 - \delta \),

\[
\left| E_{S \sim r_a^\#} [\mathbb{I}(\hat{T}_{r_a \rightarrow q^*}(S) = e_i)] - \frac{1}{n} \sum_{j=1}^{n} \mathbb{I}(\hat{T}_{r_a \rightarrow q^*}(s_{a,j}) = e_i) \right| \leq O\left( \sqrt{\frac{\ln 1/\delta}{n}} \right)
\]
by applying Lemma \ref{lem:coupling} to the (trivial) binary classification problem defined by the loss \( \ell(T, s) := \mathbb{1}(T(s) = e_i) \); clearly, the binary function class \( \{ s \mapsto \mathbb{1}(g(s) = e_i) : g \in \mathcal{G}_k \} \) cannot shatter two points (their outputs are 0 on exactly half of the line along \( v_{ij} \), for arbitrary \( j \neq i \)), thereby its Natarajan dimension (reduces to VC dimension here) is 1. The number of decision boundaries that the function \( s \mapsto \mathbb{1}(\tilde{T}_{r_a \to q}(s) = e_i) \) has is \( k \), so by the same arguments above, almost surely \( \tilde{T}_{r_a \to q} \) agrees with \( T^*_{r_a \to q} \) on all but \( k \) points from the subset of \( S_a \) that are transported to \( e_i \) (by either transport map), so

\[
\frac{1}{n} \sum_{j=1}^{n} \mathbb{1}(\tilde{T}_{r_a \to q}(s_{a,j}) = e_i) = \hat{q}(e_i) - \hat{q}(e_i) = \sum_{s \in S_a} ((\text{Id} \times \tilde{T}_{r_a \to q})\tilde{r}_a(s, e_i) - \sum_{s \in S_a} \tilde{\gamma}_a(s, e_i) \leq \sum_{s \in S_a} ((\text{Id} \times \tilde{T}_{r_a \to q})\tilde{r}_a(s, e_i) - \tilde{\gamma}_a(s, e_i)) = O(k/n),
\]

and the result follows from a union bound.

**Proof of Proposition** \[\text{Proposition} 3\] The randomized fair classifier is \( \tilde{h}_\rho(x, a) = T^*_{r_a \to q^*, \rho} \circ u \circ f(x, a) \), and its Markov kernel is given by

\[
K(s, B) = \int_{\tilde{r}_a(s) \in B} \frac{\tilde{\gamma}_a(s, B)}{\gamma_a(s, B)} d(\rho * \delta_s)(\tilde{s}),
\]

where we denote the coupling associated with \( T^*_{r_a \to q^*, \rho} \) by \( \tilde{\gamma}_a^* \).

Because \( f \) is calibrated, it is the Bayes regressor on the transformed problem defined by the joint distribution on \( (f(X, A), Y, A) \) as discussed in Section 3.2, so by Lemma \[\text{Lemma 11}\] \( \text{Err}_a(\tilde{h}_\rho) = \frac{1}{2} \int_{\Delta_k \times Y} \|s - y\|_1 d\gamma_a(s, y) \) with

\[
\gamma_a(s, y) = \int_{x \in f_a^{-1}(s)} K(x, y) d\mu^X(x) = \int_{x \in f_a^{-1}(s)} \int_{\tilde{s} \in \mathbb{R}^k} \frac{\tilde{\gamma}_a(s, y)}{\gamma_a(s, \tilde{s})} d(\rho * \delta_s)(\tilde{s}) d\mu^X(x) = \int_{\tilde{s} \in \mathbb{R}^k} \frac{\tilde{\gamma}_a(s, y)}{\tilde{r}_a(s)} d(\rho * \delta_s)(\tilde{s}),
\]

where we defined \( f_a^{-1}(s) = \{ x \in \mathcal{X} : f(x, a) = s \} \). So

\[
\text{Err}_a(\tilde{h}_\rho) = \frac{1}{2} \int_{\Delta_k \times Y} \int_{\mathbb{R}^k} \|s - y\|_1 \frac{\tilde{\gamma}_a(s, y)}{\tilde{r}_a(s)} d(\rho * \delta_s)(\tilde{s}) ds dy \leq \frac{1}{2} \int_{\Delta_k \times Y} \int_{\mathbb{R}^k} \|s - y\|_1 \frac{\tilde{\gamma}_a(s, y)}{\gamma_a(s)} d\rho(n) ds dy + \int_{\Delta_k} \int_{\mathbb{R}^k} \|n\|_1 d\rho(n) ds \leq \frac{1}{2} \left( \int_{\mathbb{R}^k} \|s - y\|_1 \frac{\tilde{\gamma}_a(s, y)}{\gamma_a(s)} d\rho(n) ds dy + \int_{\mathbb{R}^k} \|n\|_1 d\rho(n) ds \right) = \frac{1}{2} \left( \int_{\mathbb{R}^k \times Y} \|s - y\|_1 d\gamma_a(s, y) + \int_{\mathbb{R}^k} \|n\|_1 d\rho(n) ds \right)
\]

\[
= \frac{1}{2} \left( W_1(\tilde{r}_a, q^*) + \mathbb{E}_{N \sim \rho[\|N\|_1]} \right),
\]

29
where the second term on line 3 follows from a change of variable \( n := \tilde{s} - s \). Then we have
\[
\sum_{a \in [m]} \text{Err}_a(\tilde{h}_a) - \text{Err}^*_f \leq \frac{1}{2} \sum_{a \in [m]} (W_1(\tilde{r}_a, q^*) - W_1(r_a, q^*)) + \frac{m}{2} \cdot \mathbb{E}_{N \sim \rho}[\|N\|_1]
\]
\[
\leq \frac{1}{2} \sum_{a \in [m]} W_1(\tilde{r}_a, r_a) + \frac{m}{2} \cdot \mathbb{E}_{N \sim \rho}[\|N\|_1].
\]

Now, consider the coupling \( \gamma' \in \Gamma(\tilde{r}_a, r_a) \) given by \( \gamma'(\tilde{s}, s) = \rho(\tilde{s} - s)r_a(s) \), so that
\[
W_1(\tilde{r}_a, r_a) = \inf_{\gamma \in \Gamma(\tilde{r}_a, r_a)} \int \|\tilde{s} - s\|_1 \, d\gamma(\tilde{s}, s)
\]
\[
\leq \inf_{\gamma \in \Gamma(\tilde{r}_a, r_a)} \iint \|\tilde{s} - s\|_1 \rho(\tilde{s} - s)r_a(s) \, d\tilde{s} \, ds
\]
\[
\leq \inf_{\gamma \in \Gamma(\tilde{r}_a, r_a)} \iint \|(s + n) - s\|_1 \rho(n)r_a(s) \, dn \, ds
\]
\[
= \int \|n\|_1 \, d\rho(n),
\]
where line 3 follows from a change of variable \( \tilde{s} := s + n \).

On the other hand, let \( \gamma_{a, \rho} \in \Gamma(r_a, q^*) \) denote the coupling associated with the transport \( T_{r_a \to q^*} = T_{r_a \to q^*} \circ u \), then
\[
\text{Err}_a(\tilde{h}_a) = \frac{1}{2} \int \|s - y\|_1 \, d\gamma_{a, \rho}(s, y) \geq \frac{1}{2} \inf_{\gamma \in \Gamma(r_a, q^*)} \int \|s - y\|_1 \, d\gamma(s, y) = \text{Err}_a(\tilde{h}_a^*),
\]
where \( \tilde{h}_a^* \) is an optimal fair classifier derived from \( f \) achieving \( \text{Err}_f^* \).

### D Simplex-Vertex Transportation

Denote the \((k - 1)\)-simplex by \( \Delta_k = \{ x \in \mathbb{R}^k : x \geq 0, \|x\|_1 = 1 \} \), its vertices are \( \{ e_1, \cdots, e_k \} \). This section studies the simplex-vertex optimal transportation problem under the \( \ell_1 \) cost of \( c(x, y) := \|x - y\|_1 \).

We show in Lemma 16 that if the problem is semi-discrete, then there exists a Monge optimal transport from the simplex to the vertex that is unique, and is given by the \( c \)-transform of the Kantorovich potential from considering the Kantorovich-Rubinstein dual formulation of the optimal transport problem (Villani 2003). In other words, the Monge transport belongs to the following function class (with ties broken to the \( e_i \) with the smallest index \( i \)):
\[
\mathcal{G}_k := \left\{ x \mapsto e_{\text{arg min}_{e_i} \|x - e_i\|_1} : \psi \in \mathbb{R}^k \right\} \subset \{ e_1, \cdots, e_k \}^{\Delta_k},
\]
which we show in Theorem 19 admits a low function complexity of \( d_N(\mathcal{G}_k) = k - 1 \) in Natarajan dimension.

#### D.1 Geometry of Simplex-Vertex Optimal Transport

Let \( p \) be a distribution supported on \( \mathcal{X} = \Delta_k, k \geq 2 \), and \( q \) a (finite) distribution on \( \mathcal{Y} = \{ e_1, \cdots, e_k \} \), both contained in \( \mathbb{R}^k \). Consider the Kantorovich transportation problem between \( p, q \) under the \( \ell_1 \) cost,
\[
\min_{\gamma \in \Gamma(p, q)} \int_{\mathcal{X} \times \mathcal{Y}} \|x - y\|_1 \, d\gamma(x, y),
\]
Figure 4: Examples of $B_{ij}$, $C_i$, $A_i$, and the intersection of the $A_i$’s for $k = 3$. See Fig. 5 for an example where the intersection is empty when the underlying transport is not optimal.

where the min is attainable because the supports are compact.

Define the $(k-1)$-dimensional subspace (that contains $\Delta_k$)

$$\mathbb{D}^k := \{ x \in \mathbb{R}^k : \|x\|_1 = 1 \},$$

and for all $i \in [k]$ the following basis for $\mathbb{D}^k$:

$$v_{ij} := e_j - e_i, \quad j \neq i.$$ 

Also, note the following identity of the $\ell_1$ distance between a point in the simplex and its vertices (central to proving geometric properties of simplex-vertex optimal transports in the upcoming Proposition 14):

$$\forall x \in \Delta_k, \; i \in [k], \quad \|x - e_i\|_1 = 1 - x_i + \sum_{j \neq i} x_j = 1 - 2x_i + \sum_j x_j = 2(1 - x_i). \quad (10)$$

Fix a simplex-vertex transport $\gamma \in \Gamma(p,q)$. We consider its geometry properties, and use them to establish the existence of Monge transport when the problem is semi-discrete. First, we define a few items (see Fig. 4 for illustrations):

$$B_{ij} := \min \{ b \in \mathbb{R} : \gamma(\{ x \in \Delta_k : x^T v_{ij} \leq b - 1 \}, e_i) = q(e_i) \} \cup \{0\}, \quad \text{and}$$

$$C_i := \bigcap_{j \neq i} \{ x \in \Delta_k : x^T v_{ij} \leq B_{ij} - 1 \}.$$ 

If $\gamma$ is an optimal transport, it is intuitive to think that for achieving the minimum transportation cost the “boundaries” should not overstep ($B_{ij} + B_{ji} \leq 2$ for all $i \neq j$) and the “decision regions” $C_i$ of the transport should not intersect pairwise (except on a set of Lebesgue measure zero). These intuitions are indeed valid, and they follow from showing that the intersection of the following sets is nonempty

$$A_i := \bigcap_{j \neq i} \{ x \in \mathbb{D}^k : x^T v_{ij} \geq B_{ij} - 1 \}.$$ 

**Proposition 14.** If $\gamma$ is an optimal simplex-vertex transport, then $\bigcap_{i \in [k]} A_i \neq \emptyset$.

The intersection of the $A_i$’s is exactly the convex polytope considered in Section 4.2.2 for the construction of the transport map from discrete optimal transport, hence establishing its nonemptiness is not only central to proving the aforementioned geometric properties of simplex-vertex optimal
transports, but also to our procedure in Theorem 6 and its companion generalization analysis. The
proof is technical and deferred to the end of this section.

As alluded to earlier, Proposition 14 implies the existence of a mapping \( g \in G_k \) that agrees
with the decision regions, i.e., \( C_i \subseteq g^{-1}(e_i) \) for all \( i \in [k] \). In particular, given an arbitrary point
\( z \in \cap_{i \in [k]} A_i \), such a \( g \in G_k \) is given by a mapping whose decision boundaries are centered at \( z \),
since this point satisfies \( z^\top v_{ij} \leq B_{ij} - 1 \) for all \( i, j \). To get the \( \psi \in \mathbb{R}^k \) associated with this \( g \), note
that it should satisfy
\[
x^\top v_{ij} \leq z^\top v_{ij}, \forall j \neq i \iff \|x - e_i\|_1 - \psi_i \leq \|x - e_j\|_1 - \psi_j, \forall j \neq i
\]
where
\[
\|x - e_i\|_1 - \psi_i \leq \|x - e_j\|_1 - \psi_j \iff 2(1 - x_i) - \psi_i \leq 2(1 - x_j) - \psi_j
\]
\[
\iff 2(x_j - x_i) \leq \psi_i - \psi_j
\]
\[
\iff 2(x^\top v_{ij}) \leq \psi_i - \psi_j,
\]
meaning that \( \psi_i - \psi_j = 2v_{ij}^\top z = 2(z_j - z_i) \) for all \( j \neq i \), which gives a system of \( k - 1 \) equations. Set \( \psi_1 = 0 \) w.l.o.g. (because \( G_k \) is an overparameterization; the center lies on \( D^k \), a \( (k - 1) \)-dimensional subspace, so its location can be specified with \( (k - 1) \) parameters), we get
\[
\psi_i = \begin{cases} 
0 & \text{if } i = 1 \\
2(z_1 - z_i) & \text{else.}
\end{cases}
\]

An immediate consequence of the above discussions is that for any simplex-vertex optimal
transport \( \gamma \), we can extract a mapping \( g \in G_k \) that agrees with \( \gamma \) except on its boundaries:

**Corollary 15.** Let \( p, q \) be probability measures on \( \Delta_k \) and \( \{e_1, \ldots, e_k\} \), respectively, and \( c(x, y) = \|x - y\|_1 \). Let \( \gamma \in \Gamma(p, q) \) be an optimal transport between \( p, q \), then \( \exists g \in G_k \) associated with \( \psi \in \mathbb{R}^k \)

\[
\gamma(x, g(x)) = 1, \quad \forall x \in \text{supp}(p) \setminus \bigcup_{i \neq j} \{x \in D^k : \|x - e_i\|_1 - \psi_i = \|x - e_j\|_1 - \psi_j\}.
\]

When \( p \) is a continuous distribution, the set of points lying on the boundaries have measure
zero, so the transportation cost under \( g \) is the same as that of \( \gamma \), thereby showing \( g \) is a Monge
optimal transport.

**Lemma 16.** Let \( p \) be a continuous probability measure on \( X = \Delta_k \) and \( q \) a probability measure on \( \{e_1, \ldots, e_k\} \), and \( c(x, y) = \|x - y\|_1 \). Then there exists a Monge optimal transport between \( p, q \) that
is unique up to sets of measure zero w.r.t. \( p \).

**D.1.1 Programs for Point-Finding**

A point \( z \in \cap_{i \in [k]} A_i \) can be found via solving for a feasible point of the following linear program,
since each set \( A_i \) is an intersection of halfspaces:
\[
\min_{z \in \mathbb{R}^k} 0
\]
\[
s.t. \quad z^\top v_{ij} \leq B_{ij} - 1, \quad \forall i, j \in [k], i \neq j,
\]
\[
\|z\|_1 = 1.
\]
As illustrated in Fig. 3, the resulting z determines the center of the extracted transport map $g \in G_k$ and thereby its boundaries, so intuitively, mappings with larger margins should be preferred for potentially better generalizability. Therefore, we use the following quadratic program for finding $z$ that maximizes the margins to the boundaries:

$$
\max_{z \in \mathbb{R}^k} \sum_{i \neq j} ||z^T v_{ij} - (B_{ij} - 1)||^2_2 \\
\text{s.t.} \quad z^T v_{ij} \leq B_{ij} - 1, \quad \forall i, j \in [k], i \neq j, \\
||z||_1 = 1.
$$

Although ablations are not included, our internal results showed that points found using the QP are overall slightly better than those using the LP without margin maximization, in terms of both error rate and DP gap.

D.1.2 Proofs

Now we turn to the proofs.

**Proof of Lemma 16.** The existence part is a corollary to Corollary 15, so here we prove the uniqueness. Let $\gamma, \gamma' \in \Gamma(p, q)$ be two optimal transports that differ on a set with nonzero measure. We say that $e_i$ has exchanged density (with some other vertex) if there exists a set $S$ s.t. $\gamma(S, e_i) \neq \gamma'(S, e_i)$. Assume w.l.o.g. that all vertices have exchanged density; otherwise, we can remove the frivolous axes $e_i$ that do not exchange density and the densities transported to them, and end up with a lower dimensional instance where all vertices exchange density.

Let $g, g' \in G_k$ be constructed from the transports as above, and denote their centers by $z, z' \in \mathbb{D}^k$, respectively. Denote the boundaries of $g$ by $\bar{B}_{ij} = z^T v_{ij} + 1$ ($g'$ analogously), then

$$
\bar{B}_{ij} = z^T (v_{nj} - v_{ni}) + 1 = \tilde{B}_{nj} - \tilde{B}_{ni} + 1, \quad \text{for all } n, i \in [k] \text{ with } \tilde{B}_{nn} := 0.
$$

Denote the decision region of $g$ by $\bar{C}_i := \bigcap_{j \neq i} \{x \in \Delta_k : x^T v_{ij} \leq \bar{B}_{ij} - 1\}$, then $C_i \subseteq \bar{C}_i$, so $\gamma(\bar{C}_i, e_i) = \gamma(\bar{C}_i, e_i) = q(e_i)$; also, by construction, $S \subseteq C_i \iff \gamma(S, e_i) = p(S)$, because the entirety of $S$ is transported to $e_i$ under $\gamma$. Finally, define the change in $g$’s boundary by $d_{ij} := \bar{B}_{ij} - B_{ij}$, then

$$
d_{ij} = d_{nj} - d_{ni}, \quad \text{for all } n \in [k] \text{ and } i \neq j \text{ with } d_{nn} := 0.
$$

Note that at least one $\bar{C}'_i$ cannot increase in size: there must exists $d_{ij} \geq 0$ for some $i \neq j$, assume w.l.o.g. that $d_{ij} \geq d_{im}$ for all $m \neq j$, then it follows that $d_{ji} = -d_{ij} \leq 0$ and $d_{jm} = d_{im} - d_{ij} \leq 0$ for all $m \neq j$, meaning that $\bar{C}'_j \subseteq \bar{C}_j$. Combined with the assumption that $e_j$ has exchanged density, this result implies that $q(e_j) = \gamma'(\bar{C}'_j, e_j) = p(\bar{C}'_j) < p(\bar{C}_j) = q(\bar{C}_j, e_j) = q(e_j)$, which is a contradiction. 

The proof of Proposition 14 needs the following technical result, which states at a high-level that if a collection of $v_{ij}$-aligned convex sets do not intersect, then it cannot cover the entire space:

**Proposition 17.** Let $B \in \mathbb{R}^{k \times k}$ arbitrary, and define $S_i := \bigcap_{j \neq i} \{x \in \mathbb{D}^k : x^T v_{ij} \leq B_{ij} - 1\}$ for each $i \in [k]$, then $\bigcap_{i \in [k]} S_i = \emptyset \implies \bigcup_{i \in [k]} S_i \neq \mathbb{D}^k$. 

33
While this could be established with elementary arguments, for clarity, we borrow results from algebraic topology, for which interested readers are referred to [Spanier, 1981]. The concepts that we require are: if two topological spaces are homotopy equivalent then their homology groups are isomorphic, and it is clear that $\mathbb{D}^k$ is homotopy equivalent to $\mathbb{R}^{k-1}$, denoted by $\mathbb{D}^k \cong \mathbb{R}^{k-1}$, where all homology groups of $\mathbb{R}^{k-1}$ except for the zeroth homology are $\{0\}$. We also cite the following Nerve theorem:

**Theorem 18** [Bauer et al., 2022] Theorem 3.1. Let $A = \{A_1, \cdots, A_n\}$ be a finite collection of sets, and define its nerve by

$$\text{Nrv}(A) = \left\{ J \subseteq [n] : \bigcap_{i \in J} A_i \neq \emptyset \right\}.$$  

If the sets $A_i$'s are convex closed subsets of $\mathbb{R}^d$, then $\text{Nrv}(A) \cong \bigcup_{i \in [n]} A_i$.

**Proof of Proposition 17** We prove the contrapositive statement of $\bigcup_{i \in [k]} S_i = \mathbb{D}^k \implies \bigcap_{i \in [k]} S_i \neq \emptyset$ by strong induction. For the base case of $k = 2$, observe that $S_1 \cup S_2 = \{x : x^T v_{12} \leq B_{12} - 1 \text{ or } x^T v_{12} \geq 1 - B_{21}\}$, so $S_1 \cup S_2 = \mathbb{D}^2$ if and only if $B_{12} - 1 \geq 1 - B_{21}$, in which case the point $(1 - B_{12}/2, B_{12}/2) \in S_1 \cap S_2$, so the latter is not empty.

For $k > 2$, suppose $\bigcup_{i \in [k]} S_i = \mathbb{D}^k$. Our goal is to show that for all $J \subset [k]$, $\bigcap_{j \in J} S_j \neq \emptyset$. Recall that

$$S_i = \bigcap_{j \in [k], j \neq i} \{x \in \mathbb{D}^k : x^T v_{ij} \leq B_{ij} - 1\},$$  

and we define for any $J \subset [k]$ and $i \in [k]$

$$S'_{J,i} := \bigcap_{j \in J, j \neq i} \{x \in \mathbb{D}^k : x^T v_{ij} \leq B_{ij} - 1\}$$

(we drop the subscript $J$ as the discussions below will focus on a fix $J$).

We first show that $\bigcap_{i \in J} S_i \neq \emptyset$ for any $J \subset [k]$ with $|J| \leq k - 1$. By assumption, $\mathbb{D}^k = \bigcup_{i \in [k]} S_i \subset \bigcup_{i \in [k]} S'_i$, then it follows that $\bigcup_{i \in J} S'_i = \mathbb{D}^k$; otherwise we can find a line $L$ in $\mathbb{D}^k$ that is not contained in $\bigcup_{i \in J} S'_i$ in its entirety, and not contained in $\bigcup_{i \notin J} S'_i$ partially, hence contradicting $\bigcup_{i \in [k]} S'_i = \mathbb{D}^k$. Suppose not, then let $z \notin \bigcup_{i \in J} S'_i$, and consider

$$L := \left\{ z + \alpha \sum_{i \notin J, j \in J} v_{ij} : \alpha \in \mathbb{R} \right\}.$$  

No part of this line is contained in $\bigcup_{i \in J} S'_i$, because it does not contain the point $z \in L$, and $L$ runs parallel to and hence never intercepts any of the halfspaces defining each $S'_i$ for $i \in J$: let $i, j \in J$, $i \neq j$, then

$$v^T_{ij} \sum_{n \notin J, m \in J} v_{nm} = \sum_{n \notin J, m \in J} (e^T_j e_m - e^T_j e_n - e^T_i e_m + e^T_i e_n) = 1 - 0 - 1 + 0 = 0.$$  

This line is also partially not contained any $S'_i = \bigcap_{j \in J} \{x \in \mathbb{D}^k : x^T v_{ij} \leq B_{ij} - 1\}$ for $i \notin J$: let $i \notin J$ and $j \in J$, then

$$v^T_{ij} \sum_{n \notin J, m \in J} v_{nm} = \sum_{n \notin J, m \in J} (e^T_j e_m - e^T_j e_n - e^T_i e_m + e^T_i e_n) = 1 - 0 - 0 + 1 = 2;$$

34
so points on $L$ corresponding to all sufficiently large $\alpha$ are not contained in $\bigcup_{i \notin J} S_i'$.

Back to proving that $\bigcap_{i \in J} S_i \neq \emptyset$ for any $J \subset [k]$ with $|J| \leq k - 1$. Since $\bigcup_{i \notin J} S_i' = \mathbb{D}^k$, by applying the inductive hypothesis to the instance of the $|J|$-case obtained from removing the frivolous axes (namely $\{e_i : i \notin J\}$), we get $\exists z' \in \bigcap_{i \notin J} S_i'$. By the same reasoning as above, the line
$L' := \{z' + \alpha \sum_{i \notin J, j \in J} v_{ij} : \alpha \in \mathbb{R}\}$ is entirely contained in $\bigcap_{i \in J} S_i'$, and hence partially contained in $\bigcap_{i \in J} S_i = (\bigcap_{i \in J} S_i') \cap \bigcap_{i \notin J, j \in J} \{x \in \mathbb{D}^k : x^\top v_{ij} \leq B_{ij} - 1\}$, so this set is nonempty.

We have thus established that any intersection of $2, \ldots, k - 1$ sets in $\{S_i\}_{i \in [k]}$ is nonempty, and this is used to show that $\bigcap_{i \in [k]} S_i \neq \emptyset$ via the Nerve theorem. Note that the former result means that $\forall J \subset [k], 1 \leq |J| \leq k - 1, J \in \text{Nrv}(\{S_1, \ldots, S_k\})$, and we assumed in the beginning that $\bigcup_{i \in [k]} S_i = \mathbb{D}^k$, then it must be that $[k] \in \text{Nrv}(\{S_1, \ldots, S_k\})$ as well; otherwise, the nerve represents a $(k - 1)$-dimensional simplex (each $n$-face is represented by its $n - 1$ vertices) without its interior (represented by $[k]$), whose $(k - 2)$-th homology is $\mathbb{Z}$ (Spanier 1981), so $\bigcup_{i \in [k]} S_i \neq \mathbb{D}^k$ by Theorem 18 which is a contradiction. Hence the nerve contains $[k]$, meaning that $\bigcap_{i \in [k]} S_i \neq \emptyset$.

**Proof of Proposition 14** Suppose $\bigcap_{i \in [k]} A_i = \emptyset$, then $\exists z \in \bigcap_{i \in [k]} (\mathbb{D}^k \setminus A_i)$ by Proposition 17. It follows by construction that there exists a mapping $\mathbf{u} : [k] \to [k], \mathbf{u}(i) \neq i$ s.t. $\forall i \in [k], z^\top v_{\mathbf{u}(i)i} < B_{\mathbf{u}(i)i} - 1$. Thereby, there exist $m > 0$ and measurable sets $F_i \subset \{x : x^\top v_{\mathbf{u}(i)i} > z^\top v_{\mathbf{u}(i)i}\} \subset C_i$ s.t. $\gamma(F_i, e_i) := m_{i} \geq m$. Note that there exists a nonempty $J \subseteq [m]$ s.t. the set of undirected edges $\{(i, \mathbf{u}(i)) : i \in J\}$ forms a cycle.

Observe that the coupling $\gamma' \in \Gamma(p, q)$ given by

$$\gamma'(B, e_i) = \begin{cases} \gamma(B, e_i) & \text{if } i \notin J, \\
\gamma(B \cap (\Delta_k \setminus F_i), e_i) + \frac{m_i - m}{m_i} \gamma(B \cap F_i, e_i) + \frac{m}{m_{\mathbf{u}^{-1}(i)}} \gamma(B \cap F^{-1}_{\mathbf{u}^{-1}(i)}, e_{\mathbf{u}^{-1}(i)}) & \text{else} \end{cases}$$

Figure 5: Illustration of the construction in proof of Proposition 14 for $k = 3$. The $A_i$'s here correspond to the $S_i$'s in the statement of Proposition 17.
has a lower transportation cost than $\gamma$ (see Fig. 5 for an illustration), because

$$\int_{\Delta_k \times \mathcal{Y}} \|x - y\|_1 \ d(\gamma - \gamma')(x, y) = \sum_{i \in J} \frac{m}{m_i} \int_{F_i} \left(\|x - e_i\|_1 - \|x - e_{u(i)}\|_1\right) d\gamma(x, e_i)
$$

$$= \sum_{i \in J} \frac{2m}{m_i} \int_{F_i} (x_{u(i)} - x_i) d\gamma(x, e_i)
$$

$$= \sum_{i \in J} \frac{2m}{m_i} \int_{F_i} x^\top v_{iu(i)} d\gamma(x, e_i)
$$

$$> \sum_{i \in J} \frac{2m}{m_i} \int_{F_i} z^\top v_{iu(i)} d\gamma(x, e_i)
$$

$$= 2m \sum_{i \in J} z^\top v_{iu(i)} = 2m \sum_{i \in J} (z_{u(i)} - z_i) = 0,$$

where line 2 follows from Eq. (10).

D.2 Natarajan Dimension of $\mathcal{G}_k$

**Theorem 19.** The Natarajan dimension of the the function class $\mathcal{G}$ is $k - 1$.

Before proving this result, we need a lemma.

**Lemma 20.** For $k \geq 2$, consider the set of vectors

$$M \subseteq \hat{M} = \{-e_i + e_j \mid i \neq j \in [k]\} \subseteq \mathbb{R}^k.$$

If $|M| \geq k$, then $M$ is NOT linearly independent over $\{-1, 0, 1\}$.

**Proof.** When $k = 2$, $M = \hat{M} = \{-e_1 + e_2, -e_2 + e_1\}$; hence the statement is clearly true. Now suppose the statement is true for all $t \in \mathbb{N}$, and now re-write

$$M \subseteq \hat{M} = \{-e_i + e_j \mid i \neq j \in [t + 1]\} \subseteq \mathbb{R}^{t+1}$$

and assume $|M| = t + 1$. By the pigeonhole principle, there must exist $i \in [k]$ such that $e_i$ appears at most twice in $M$.

If such $i$ never appears, then we may regard $M$ and $\hat{M}$ to be a subspace of $\mathbb{R}^t$. By the inductive hypothesis, the statement is true.

If such $i$ appears only once in, without loss of generality, $v = -e_i + e_j$ for some $j \neq i$, then by eliminating this vector, the remaining set $M' = M \setminus \{v\}$ has $t$ elements as a subset of $\mathbb{R}^t$ or $\mathbb{R}^{t-1}$; by the inductive hypothesis, $M'$ is linearly dependent over $\{-1, 0, 1\}$.

If such $i$ appears exactly twice in $v = -e_i + e_j$ and $w = -e_i + e_k$, then clearly the set

$$M^* = (M \setminus \{v, w\}) \cup \{-v + w = -e_j + e_k\}$$

has $t$ elements and is a subset of $\mathbb{R}^t$. Again, by the inductive hypothesis, the statement is true. 

**Proof of Theorem 19.** We divide this proof into two parts: the first part proves $d_N(\mathcal{G}) < k$, and the second part proves $d_N(\mathcal{G}) \geq k - 1$. 

36
Part 1. \( d_{\mathcal{N}}(\mathcal{G}) < k \).

We first prove \( d_{\mathcal{N}}(\mathcal{G}) < k \). Let \( C = \{x_1, \ldots, x_k\} \subseteq \Delta_k \), and let \( f_0, f_1 : C \to [k] \) be any two mappings such that for any \( x \in C \), \( f_0(x) \neq f_1(x) \).

Note that, for every \( x_i \in C \) and the induced \( (f_0(x_i), f_1(x_i)) \in [k] \times [k] \), we may consider it as a hyperplane \( H_i \) with the normal vector \(-e_{f_0(x_i)} + e_{f_1(x_i)}\):

\[
H_i = \left\{ y \in \mathbb{R}^k : -y_{f_0(x_i)} + y_{f_1(x_i)} = \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1 \right\}.
\]

Indeed, if there exists \( \psi \in \mathbb{R}^k \) such that \( g_{\psi}(x_i) = f_0(x_i) \), where \( g_{\psi} \in \mathcal{G}_k \), then:

\[-\psi_{f_0(x_i)} + \psi_{f_1(x_i)} < \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1;\]

similarly, if \( g_{\psi}(x_i) = f_1(x_i) \), then:

\[-\psi_{f_0(x_i)} + \psi_{f_1(x_i)} > \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1.\]

Since we have \( k \) points, we have produced \( k \) hyperplanes whose normal vectors form the following set

\[
M = \left\{ v_i = -e_{f_0(x_i)} + e_{f_1(x_i)} \mid i \in [k + 1] \right\}.
\]

If \( |M| < k \), i.e., for some \( i \neq j \in [k] \), \( f_0(x_i) = f_0(x_j) \) where \( b \in [2] \). Without loss of generality, assume \( f_0(x_i) = f_0(x_j) < f_1(x_i) = f_1(x_j) \).

Again without loss of generality, assume

\[
\|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1 \leq \|x_j - e_{f_1(x_j)}\|_1 - \|x_j - e_{f_0(x_j)}\|_1.
\]

Consider the partition \( C = B \cup (C - B) \), where \( B \ni x_i \) but \( B \) does not contain \( x_j \). If \( \mathcal{G} \) shatters \( C \), then we have

\[-\psi_{f_0(x_i)} + \psi_{f_1(x_i)} \leq \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1.\]

On the other hand, because \( x_j \) is not in \( B \), we must also have:

\[-\psi_{f_0(x_j)} + \psi_{f_1(x_j)} > \|x_j - e_{f_1(x_j)}\|_1 - \|x_j - e_{f_0(x_j)}\|_1 \geq \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1,
\]

which is a contradiction.

Now suppose \( M \) has no repetitive elements, i.e., \( |M| = k \); yet, \( M \subseteq \mathbb{R}^k \). Therefore, by Lemma [2], there must exist a \( v_i \) and \( r \leq k \) vectors \( v_{q_1}, \ldots, v_{q_r} \) such that

\[
v_i = \sum_{m=1}^{r} a_{q_m} \cdot v_{q_m}, \quad a_{q_m} \in \{-1, 1\}. \tag{11}
\]

Without loss of generality, we may assume \( f_0(x_i) < f_1(x_i) , a_{q_1} = \cdots = a_{q_d} = 1 \) and \( a_{q_{d+1}} = \cdots = a_{q_r} = -1 \). Consider the partition of \( C = B \cup (C - B) \), where \( B \) contains \( \{v_{q_1}, \ldots, v_{q_d}\} \) but does not contain \( \{v_{q_{d+1}}, \ldots, v_{q_r}, v_i\} \). Moreover, we assume there exists \( \psi \in \mathbb{R}^k \) such that \( g_{\psi}(x) = f_0(x) \) for \( x \in B \) and \( g_{\psi}(x) = f_1(x) \) for \( x \notin B \). Following the argument above, we obtain the following
inequalities:

\[-\psi_{f_0(x_{q_1})} + \psi_{f_1(x_{q_1})} \leq \|x_{q_1} - e_{f_1(x_{q_1})}\|_1 - \|x_{q_1} - e_{f_0(x_{q_1})}\|_1;\]

\[\vdots\]

\[-\psi_{f_0(x_{q_d})} + \psi_{f_1(x_{q_d})} \leq \|x_{q_d} - e_{f_1(x_{q_d})}\|_1 - \|x_{q_d} - e_{f_0(x_{q_d})}\|_1;\]

\[-\psi_{f_0(x_{q_{d+1}})} + \psi_{f_1(x_{q_{d+1}})} \geq \|x_{q_{d+1}} - e_{f_1(x_{q_{d+1}})}\|_1 - \|x_{q_{d+1}} - e_{f_0(x_{q_{d+1}})}\|_1;\]

\[\vdots\]

\[-\psi_{f_0(x_{q_r})} + \psi_{f_1(x_{q_r})} \leq \|x_{q_r} - e_{f_1(x_{q_r})}\|_1 - \|x_{q_r} - e_{f_0(x_{q_r})}\|_1;\]

\[-\psi_{f_0(x_i)} + \psi_{f_1(x_i)} > \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1.\]

By reversing the order of the inequalities from the \((d + 1)\)-th to the \(r\)-th, we obtain the following system:

\[-\psi_{f_0(x_{q_1})} + \psi_{f_1(x_{q_1})} \leq \|x_{q_1} - e_{f_1(x_{q_1})}\|_1 - \|x_{q_1} - e_{f_0(x_{q_1})}\|_1;\]

\[\vdots\]

\[-\psi_{f_0(x_{q_d})} + \psi_{f_1(x_{q_d})} \leq \|x_{q_d} - e_{f_1(x_{q_d})}\|_1 - \|x_{q_d} - e_{f_0(x_{q_d})}\|_1;\]

\[-\psi_{f_0(x_{q_{d+1}})} + \psi_{f_1(x_{q_{d+1}})} \leq \|x_{q_{d+1}} - e_{f_0(x_{q_{d+1}})}\|_1 - \|x_{q_{d+1}} - e_{f_1(x_{q_{d+1}})}\|_1;\]

\[\vdots\]

\[-\psi_{f_0(x_{q_r})} + \psi_{f_0(x_{q_r})} \leq \|x_{q_r} - e_{f_0(x_{q_r})}\|_1 - \|x_{q_r} - e_{f_1(x_{q_r})}\|_1;\]

\[-\psi_{f_0(x_i)} + \psi_{f_1(x_i)} > \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1.\]

Be aware that, because of Eq. \([11]\), we have

\[
\|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1 < -\psi_{f_0(x_i)} + \psi_{f_1(x_i)}
\]

\[
= \sum_{t=1}^{d} \left( -\psi_{f_0(x_{q_t})} + \psi_{f_1(x_{q_t})} \right) + \sum_{t=d+1}^{r} \left( -\psi_{f_1(x_{q_t})} + \psi_{f_0(x_{q_t})} \right)
\]

\[
\leq \sum_{t=1}^{d} \left( \|x_{q_t} - e_{f_1(x_{q_t})}\|_1 - \|x_{q_t} - e_{f_0(x_{q_t})}\|_1 \right) + \sum_{t=d+1}^{r} \left( \|x_{q_t} - e_{f_0(x_{q_t})}\|_1 - \|x_{q_t} - e_{f_1(x_{q_t})}\|_1 \right).
\]

If \(G\) shatters \(C\), then for the partition of \(C = B' \cup (C - B')\) where \(B'\) contains \(\{v_{q_{d+1}}, \ldots, v_{q_r}, v_i\}\) but does not contain \(\{v_{q_1}, \ldots, v_{q_d}\}\), there must exist \(\pi \in \mathbb{R}^k\) such that \(g_{\pi}(x) = f_0(x)\) for \(x \in B'\) and \(g_{\pi}(x) = f_1(x)\) for \(x \notin B'\). Following this fact, we obtain the following system:

\[-\pi_{f_0(x_{q_1})} + \pi_{f_1(x_{q_1})} \geq \|x_{q_1} - e_{f_1(x_{q_1})}\|_1 - \|x_{q_1} - e_{f_0(x_{q_1})}\|_1;\]

\[\vdots\]

\[-\pi_{f_0(x_{q_d})} + \pi_{f_1(x_{q_d})} \geq \|x_{q_d} - e_{f_1(x_{q_d})}\|_1 - \|x_{q_d} - e_{f_0(x_{q_d})}\|_1;\]

\[-\pi_{f_0(x_{q_{d+1}})} + \pi_{f_1(x_{q_{d+1}})} \leq \|x_{q_{d+1}} - e_{f_1(x_{q_{d+1}})}\|_1 - \|x_{q_{d+1}} - e_{f_0(x_{q_{d+1}})}\|_1;\]

\[\vdots\]

\[-\pi_{f_0(x_{q_r})} + \pi_{f_1(x_{q_r})} \leq \|x_{q_r} - e_{f_1(x_{q_r})}\|_1 - \|x_{q_r} - e_{f_0(x_{q_r})}\|_1;\]

\[-\pi_{f_0(x_i)} + \pi_{f_1(x_i)} \leq \|x_i - e_{f_1(x_i)}\|_1 - \|x_i - e_{f_0(x_i)}\|_1.\]
Again, because of Eq. (11), we have

\[ -\pi_{f_0}(x_{q_1}) + \pi_{f_1}(x_{q_1}) \geq \|x_{q_1} - e_{f_1}(x_{q_1})\|_1 - \|x_{q_1} - e_{f_0}(x_{q_1})\|_1; \]

\[ \vdots \]

\[ -\pi_{f_0}(x_{q_d}) + \pi_{f_1}(x_{q_d}) \geq \|x_{q_d} - e_{f_1}(x_{q_d})\|_1 - \|x_{q_d} - e_{f_0}(x_{q_d})\|_1; \]

\[ -\pi_{f_1}(x_{q_{d+1}}) + \pi_{f_0}(x_{q_{d+1}}) \geq \|x_{q_{d+1}} - e_{f_0}(x_{q_{d+1}})\|_1 - \|x_{q_{d+1}} - e_{f_1}(x_{q_{d+1}})\|_1; \]

\[ \vdots \]

\[ -\pi_{f_1}(x_{q_r}) + \pi_{f_0}(x_{q_r}) \geq \|x_{q_r} - e_{f_0}(x_{q_r})\|_1 - \|x_{q_r} - e_{f_1}(x_{q_r})\|_1; \]

\[ -\pi_{f_0}(x_i) + \pi_{f_1}(x_i) \leq \|x_i - e_{f_1}(x_i)\|_1 - \|x_i - e_{f_0}(x_i)\|_1. \]

Again, because of Eq. (11), we have

\[
\sum_{t=1}^{d} \left( \|x_{q_t} - e_{f_1}(x_{q_t})\|_1 - \|x_{q_t} - e_{f_0}(x_{q_t})\|_1 \right) + \sum_{t=d+1}^{r} \left( \|x_{q_t} - e_{f_0}(x_{q_t})\|_1 - \|x_{q_t} - e_{f_1}(x_{q_t})\|_1 \right)
\leq \sum_{t=1}^{d} \left( -\pi_{f_0}(x_{q_t}) + \pi_{f_1}(x_{q_t}) \right) + \sum_{t=d+1}^{r} \left( -\pi_{f_1}(x_{q_t}) + \pi_{f_0}(x_{q_t}) \right)
= -\pi_{f_0}(x_i) + \pi_{f_1}(x_i)
\leq \|x_i - e_{f_1}(x_i)\|_1 - \|x_i - e_{f_0}(x_i)\|_1.
\]

The inequality above contradicts Equations 8-11. Therefore, the second shattering condition is never satisfied for any set of cardinality at least \( k \). As a result, \( d_N(\mathcal{G}) < k \).

**Part 2.** \( d_N(\mathcal{G}) \geq k - 1 \).

Now we prove \( d_N(\mathcal{G}_k) \geq k - 1 \) by showing that, for every \( k \), \( \mathcal{G}_k \) shatters at least one set \( C_{k-1} \) of cardinality \( k - 1 \) inside \( \Delta_k \). When \( k = 2 \), we are exactly studying the \( d_{VC}(\mathcal{G}_2) \). To show it is one, simply assume a point \( x \) is labelled as \( i \in [2] \) and let \( \psi = 2e_i \). Then clearly, if \( j \neq i \),

\[
\|a - e_i\|_1 - \psi_i = \|a - e_i\| - 2 < 0 \leq \|a - e_j\| - 1 - \psi_j.
\]

Now suppose \( d_k(\mathcal{G}_k) \geq k - 1 \) for any \( k \in \mathbb{N} \), so \( \mathcal{G}_k \) shatters a set \( C_{k-1} = \{x'_1, \ldots, x'_{k-1}\} \) with mappings \( f'_0 \) and \( f'_1 : C_{k-1} \to [k] \) satisfying the two conditions of shattering. Thus, for every partition \( C_{k-1} = B_{k-1} \cup (C_{k-1} - B_{k-1}) \) where \( B_{k-1} = \{x'_{q_1}, \ldots, x'_{q_d}\} \), there exists a \( \psi' \in \mathbb{R}^k \) that satisfies the following system of \( k - 1 \) inequalities:

\[
-\psi'_0(x'_{q_1}) + \psi'_1(x'_{q_1}) \leq \|x'_{q_1} - e_{f'_1}(x'_{q_1})\|_1 - \|x'_{q_1} - e_{f'_0}(x'_{q_1})\|_1
\]

\[ \vdots \]

\[
-\psi'_0(x'_{q_d}) + \psi'_1(x'_{q_d}) \leq \|x'_{q_d} - e_{f'_1}(x'_{q_d})\|_1 - \|x'_{q_d} - e_{f'_0}(x'_{q_d})\|_1
\]

\[
-\psi'_0(x'_{q_{d+1}}) + \psi'_1(x'_{q_{d+1}}) \geq \|x'_{q_{d+1}} - e_{f'_1}(x'_{q_{d+1}})\|_1 - \|x'_{q_{d+1}} - e_{f'_0}(x'_{q_{d+1}})\|_1
\]

\[ \vdots \]

\[
-\psi'_0(x'_{q_{k-1}}) + \psi'_1(x'_{q_{k-1}}) \geq \|x'_{q_{k-1}} - e_{f'_1}(x'_{q_{k-1}})\|_1 - \|x'_{q_{k-1}} - e_{f'_0}(x'_{q_{k-1}})\|_1.
\]
the notation \( e_k \) emphasizes its dimension \( k - 1 \). Note that if \( \text{image}(f'_0) \cup \text{image}(f'_1) \) is a proper subset of \([k]\), then the "free" variables in \( \psi' \) can be any real real number.

Now consider the class \( \mathcal{G}_{k+1} \) and the set \( C_k = \{ x_1, \ldots, x_{k-1}, x_k \} \), where \( x_i = (x_i', \ldots, x_i', 0) \) for \( i \in [k-1] \), and \( x_k = e_{k+1} \). Define the two mappings \( f_0, f_1 : C_k \to [k+1] \), where \( f_0(x_i) = f_0(x_i') \), \( f_1(x_i) = f_1(x_i') \) \( \forall i \in [k-1] \), and \( f_0(x_k) = 1, f_1(x_k) = k + 1 \). Observe that, for every \( i \in [k] \) and \( b \in [2] \), we have

\[
\| x_i - e_{f_0(x_i)}^{k+1} \|_1 = \| x_i - e_{f_1(x_i)}^{k+1} \|_1. \tag{12}
\]

Consider the vector \( \psi = (\psi_1', \ldots, \psi_k, \psi_{k+1}) \in \mathbb{R}^{k+1} \), where the value of \( \psi_{k+1} \) is undetermined at this moment. Because of Eq. \( (12) \), \( \psi \) satisfies the following system of \( k - 1 \) inequalities:

\[
- \psi f_0(x_{q_1}) + \psi f_1(x_{q_1}) \leq \| x_{q_1} - e_{f_1(x_{q_1})}^{k+1} \|_1 - \| x_{q_1} - e_{f_0(x_{q_1})}^{k+1} \|_1 \\
\vdots \\
- \psi f_0(x_{q_d}) + \psi f_1(x_{q_d}) \leq \| x_{q_d} - e_{f_1(x_{q_d})}^{k+1} \|_1 - \| x_{q_d} - e_{f_0(x_{q_d})}^{k+1} \|_1 \\
- \psi f_0(x_{q_{d+1}}) + \psi f_1(x_{q_{d+1}}) \geq \| x_{q_{d+1}} - e_{f_1(x_{q_{d+1}})}^{k+1} \|_1 - \| x_{q_{d+1}} - e_{f_0(x_{q_{d+1}})}^{k+1} \|_1 \\
\vdots \\
- \psi f_0(x_{q_{k-1}}) + \psi f_1(x_{q_{k-1}}) \geq \| x_{q_{k-1}} - e_{f_1(x_{q_{k-1}})}^{k+1} \|_1 - \| x_{q_{k-1}} - e_{f_0(x_{q_{k-1}})}^{k+1} \|_1.
\]

Consider the partition \( C_k = B_k \cup (C_k - B_k) \); without loss of generality, we may assume \( B_k = \{ x_{q_1}, \ldots, x_{q_d}, x_k \} \). For \( \mathcal{G}_{k+1} \) to shatter \( C_k \) with respect to this partition, \( \psi \) must satisfy one more inequality:

\[
- \psi f_0(x_k) + \psi f_1(x_k) = -\psi_1 + \psi_{k+1} \\
= \| x_k - e_{f_1}^{k+1} \|_1 - \| x_k - e_{f_0}^{k+1} \|_1 \\
= \| e_{k+1}^{k+1} - e_{f_0}^{k+1} \|_1 - \| e_{k+1}^{k+1} - e_{f_1}^{k+1} \|_1 \\
= -2;
\]

the third line follows because \( x_k = e_{k+1} \). Note that, \( \psi_1 \) is given, and \( \psi_{k+1} \in \mathbb{R} \) only has to satisfy one inequality, so we can easily pick a value. If \( x_k \notin B_k \), we must have \(-\psi_1 + \psi_{k+1} > -2 \) and again we have infinitely many choices. Therefore, the second condition of shattering is satisfied for this partition.

Note that, since the partition of \( C_{k-1} \) for the \( k \)-dimensional case is arbitrary, the result also applies for any arbitrary partition for the \((k + 1)\)-dimensional case, and we just showed that the membership of \( x_k \) does not matter. Therefore, \( \mathcal{G}_{k+1} \) shatters \( C_k \) and concludes the proof. ■

### E Experiment Details

Statistics of the datasets are included in Tables \( 4 \) to \( 8 \) where the numbers indicate the proportion (%) of the classes in each group. We also indicate the DP gap on the datasets, and compute the maximum fair empirical accuracy, interpreted as the maximum attainable accuracy on each dataset (not population) by the perfect classifier satisfying DP. This comes from solving the TV-barycenter problem in Eq. \( (7) \) using statistics of the group-wise class distributions of the datasets. The differences between the maximum empirical fair accuracies and 100 (perfect accuracy on the datasets) serve as (approximated) upper bounds on the minimum costs of DP fairness, to which the costs of fairness in our experiments are observed to be correlated (Table \( 2 \)).

For simplex projection, we use the implementation by [Blondel et al. 2014].
Table 3: Results of pre-trained predictors (pre.) and post-processed classifiers (post.) evaluated on test set. Group-balanced accuracy is computed from the balanced error rate in Eq. (1), and DP gap is defined in Eq. (3). Post-processing is performed with Laplace(0, 0.2/|Y| · I) smoothing. Cf. Table 2.

| Dataset | Sensitive Attribute | | | Group-Balanced Acc. | | | DP Gap |
|---------|---------------------|---|---|---------------------|---|---|---------------------|
|         |                     | | | Pre. | Post. | ∆ | Pre. | Post. | ∆ |
| Adult   | Gender              | 2 | 2 | 86.35 | 83.42 | -3.39% | 0.1556 | 0.0007 | -99.95% |
|         | Race                | 5 | 2 | 86.97 | 85.09 | -2.16% | 0.1664 | 0.0287 | -82.75% |
|         | G. & R.             | 10 | 2 | 88.22 | 85.39 | -3.21% | 0.2751 | 0.0670 | -75.66% |
| BiosBias| Gender              | 2 | 28 | 85.86 | 80.61 | -6.11% | 0.2861 | 0.1280 | -55.26% |
| Communities | Race | 4 | 5 | 67.36 | 60.83 | -9.70% | 0.5321 | 0.1191 | -77.61% |

**BERT Hyperparameters.** On BiosBias, we fine-tune a neural language model from the BERT Base (uncased) checkpoint ([Devlin et al., 2019](#)) for three epochs with the AdamW optimizer on squared loss with the one-hot labels. We use a batch size of 32, learning rate of 2e-5, linear decay learning rate schedule with a warmup ratio of 0.1, $L^2$ weight decay rate of 0.01 (equivalent to $L^2$ regularization with strength 0.01 times the learning rate), and we clip the norm of the gradients to 1. The PyTorch implementation is included in our provided code, built with the Hugging Face Transformers library ([Wolf et al., 2020](#)).

**Smoothing Experiments.** Post-processing results with Laplace(0, 0.2/|Y| · I) smoothing are included in Table 3. During post-processing, for each input sample from the dataset we generate 10 points (20 on Adult dataset) via random perturbations, and during inference the results are averaged over 1000 random perturbations of each input. Note that in our smoothing experiments on Adult, DP gap is sometimes not reduced compared to Table 2. This is attributed to the variance from the random noise; further increasing the number of perturbations during post-processing should yield more consistent results from lowered variance.
Table 4: Adult dataset statistics, grouped by gender.  
DP gap is 0.1945, and maximum empirical fair accuracy is 90.27.

| Class \ Group | Female | Male |
|---------------|--------|------|
| ≤ 50k         | 89.07  | 69.62|
| > 50k         | 10.93  | 30.38|
| Count         | 14423  | 22732|

Table 5: Adult dataset statistics, grouped by race.  
DP gap is 0.1522, and maximum empirical fair accuracy is 94.29.

| Class \ Group | American Indian & Eskimo | Asian & Pacific Islander | Black | Other | White |
|---------------|--------------------------|--------------------------|-------|-------|------|
| ≤ 50k         | 88.30                    | 73.07                    | 87.92 | 87.68 | 74.60|
| > 50k         | 11.70                    | 26.93                    | 12.08 | 12.32 | 25.40|
| Count         | 470                      | 1519                     | 4685  | 406   | 41762|

Table 6: Adult dataset statistics, grouped by gender & race.  
DP gap is 0.2821, and maximum empirical fair accuracy is 93.28.

| Class \ Group \ Gender | Accountant | Architect | Attorney | Chiropractor | Comedian | Composer |
|-------------------------|------------|-----------|----------|--------------|----------|----------|
| Female                  | 1.14       | 1.32      | 6.86     | 0.38         | 0.33     | 0.50     |
| Male                    | 1.69       | 3.65      | 9.52     | 0.90         | 1.04     | 2.22     |
| Count                   | 185        | 2308      | 155      | 13027        |          |          |

Table 7: Biosbias dataset statistics, grouped by gender.  
DP gap is 0.2312, and maximum empirical fair accuracy is 88.44.

| Group \ Class \ Class | Accountant | Architect | Attorney | Chiropractor | Comedian | Composer |
|------------------------|------------|-----------|----------|--------------|----------|----------|
| Female                 | 1.14       | 1.32      | 6.86     | 0.38         | 0.33     | 0.50     |
| Male                   | 1.69       | 3.65      | 9.52     | 0.90         | 1.04     | 2.22     |
| Count                  | 185        | 2308      | 155      | 13027        |          |          |

Table 8: Communities dataset statistics, grouped by race.  
DP gap is 0.5815, and maximum empirical fair accuracy is 76.68.

| Class \ Group | Asian | Black | Hispanic | White |
|---------------|------|-------|----------|-------|
| [0.0, 0.2]    | 76.87| 34.06 | 41.73    | 92.21 |
| [0.2, 0.4]    | 15.66| 27.97 | 36.75    | 6.08  |
| [0.4, 0.6]    | 3.20 | 18.59 | 13.12    | 1.22  |
| [0.6, 0.8]    | 2.85 | 9.22  | 4.99     | 0.24  |
| [0.8, 1.0]    | 1.42 | 10.16 | 3.41     | 0.24  |
| Count         | 562  | 640   | 381      | 411   |