Fair and Optimal Classification via Post-Processing Predictors*

Ruicheng Xian[1] Lang Yin[1] Han Zhao[1]

February 15, 2023

Abstract

To address the bias exhibited by machine learning models, fairness criteria impose statistical constraints to ensure equal treatment to all demographic groups, but typically at a cost to model performance. Understanding this tradeoff, therefore, underlies the design of fair and effective algorithms. This paper completes the characterization of the inherent tradeoff of demographic parity on classification problems in the most general multigroup, multiclass, and noisy setting. Specifically, we show that the minimum error rate achievable by randomized and attribute-aware classifiers is given by the optimal value of a Wasserstein-barycenter problem. More practically, this reformulation leads to a simple procedure for post-processing any pre-trained predictors to satisfy demographic parity in the general setting, which, in particular, yields the optimal fair classifier when applied to the Bayes optimal predictor. We provide suboptimality and finite sample analyses for our procedure, and demonstrate precise control of the tradeoff of error rate for fairness on real-world datasets provided sufficient data.

1 Introduction

Machine learning models trained on biased data are observed to propagate and exacerbate the bias against historically underrepresented and disadvantaged demographic groups at inference time (Barocas and Selbst, 2016; Berk et al., 2021). This has prompted studies on aspects of fairness concerning their usage, especially as they expand to high-stakes domains such as criminal justice, healthcare, and finance (Berk et al., 2021). To mitigate the potential bias, a variety of fairness criteria and algorithms have been proposed (Barocas et al., 2019; Caton and Haas, 2020), which impose mathematical or statistical constraints on the model to ensure equal treatment under the respective notion of fairness. Typically, these algorithms induce a cost to model performance as they improve model fairness (Calders et al., 2009; Corbett-Davies et al., 2017).

It is often unclear, however, whether the degraded performance is attributed to artifacts of the algorithm, or possibly to the *inherent tradeoff*—predictive power that must be given up for satisfying the criteria (Hardt et al., 2016; Zhao and Gordon, 2022). Hence, the design of fair and effective algorithms necessitates the understanding of this tradeoff, which would also provide insight to the implications of fairness in machine learning. Yet, it remains open for most fairness criteria and problem settings.

For the group fairness criterion of *demographic parity* (DP; Definition 2.1), a.k.a. statistical parity, which requires statistical independence between model output and demographic group membership (Calders et al., 2009), Le Gouic et al. (2020) and Chzhen et al. (2020) characterized the tradeoff of mean squared error (MSE) for fairness on regression problems. On classification problems, the inherent tradeoff in terms of error rate has only been studied under special cases: Denis et al. (2022)

*Comparison to *arXiv:2211.01528*v2: fixed bug in Theorem 4.4.

[1]University of Illinois Urbana-Champaign. {rxian2,langyin2,hanzhao}@illinois.edu.

assumed binary groups, Zeng et al. (2022a) and Gaucher et al. (2022) assumed binary class labels, and Zhao and Gordon (2022) assumed that the data distribution is *noiseless*, i.e., the (unconstrained) Bayes error rate is zero. Our work closes this gap and completes the characterization of the tradeoff of DP fairness in the most general classification setting.

Contributions. This paper considers learning *randomized* and *attribute-aware* classifiers under DP fairness (allowing relaxations) in the general setting of multigroup, multiclass, and potentially *noisy* data distributions with nonzero (unconstrained) Bayes error rate. We show that:

1. The minimum classification error rate under DP is given by the optimal value of a (relaxed) Wasserstein-barycenter problem (Section 3.1).

2. This reformulation reveals that the *optimal fair classifier*, one that satisfies DP while achieving the minimum error, is composed of the (unconstrained) Bayes optimal predictor (minimum MSE regressor of the one-hot labels) and optimal transports, which arise from solving the barycenter problem (Section 3.2).

3. Based on the decomposition, we propose a post-processing procedure that is applicable to any pre-trained predictors for DP fairness (Section 3.3). Our procedure is instantiated to operate on finite samples of unlabeled data (Section 4).

   To our knowledge, this is the first DP post-processing procedure for the general classification setting with suboptimality and finite sample analyses.[2]

4. Experiments on real-world datasets demonstrate the effectiveness of our procedure (Section 5). Provided sufficient training data, it achieves precise control of the tradeoff during inference.

1.1 Related Work

Inherent Tradeoff. Many analyses of the tradeoff of DP fairness leverage the concept of barycenter (reviewed below), because of the analogy to the independence constraint. Intuitively, by treating the barycenter as the model output distribution that is required to be identical across groups, and setting the quantities over which the barycenter is computed to the optimal unconstrained (output) distributions on each group, the sum of distances to the barycenter naturally emerges as the minimum error under DP.

We review such characterizations established in prior work in Table 1, and compare them to ours. Let the input be denoted by X, group membership by A, and target variable by Y (set to the one-hot label on classification problems). Let r_a^* denote the distribution of the conditional mean, $\mathbb{E}[Y \mid X, A = a]$, i.e., the minimum MSE estimates of Y given X on each group (for classification, these are the class probabilities). Then under the strict DP constraint, for group-balanced error, on

- regression problems with MSE (Le Gouic et al., 2020; Chzhen et al., 2020), the minimum *excess* risk is given by the Wasserstein-2-barycenter (under the ℓ_2 metric) over the r_a^*'s: Eq. (1);

- noiseless classification problems (Zhao and Gordon, 2022), the minimum (excess) error rate is given by the TV-barycenter over the class marginals, $p_a^*(e_i) \coloneqq \mathbb{P}(Y = e_i \mid A = a)$: Eq. (2), where D_{TV} denotes the total variation (TV) distance;

- classification in the general setting (ours), the minimum error rate is given by the Wasserstein-1 barycenter (under the ℓ_1 metric): Eq. (3).

[2]Our code is available at `https://github.com/rxian/fair-classification`.

Table 1: Characterizations of the inherent tradeoff of (strict) DP fairness.

Problem Setting	Minimum Achievable Risk	
Regression with MSE (Le Gouic et al., 2020; Chzhen et al., 2020)	excess risk = $\min_{q:\mathrm{supp}(q)\subseteq\mathbb{R}} \frac{1}{\lvert\mathcal{A}\rvert}\sum_{a\in\mathcal{A}} W_2^2(r_a^*, q)$	(1)
Noiseless Classification (Zhao and Gordon, 2022)	excess = min. risk = $\min_{q:\mathrm{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \frac{1}{\lvert\mathcal{A}\rvert}\sum_{a\in\mathcal{A}} D_{\mathrm{TV}}(p_a^*, q)$	(2)
Classification (Theorem 3.2)	minimum risk = $\min_{q:\mathrm{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \frac{1}{\lvert\mathcal{A}\rvert}\sum_{a\in\mathcal{A}} \frac{1}{2}W_1(r_a^*, q)$	(3)

Note that, first, the barycenter in Eq. (3) is restricted to the vertices of the simplex, which correspond to the one-hot labels. Combined with the fact that the error rate is the expected $\frac{1}{2}\ell_1$ distance between the true class probabilities and the output labels, the sum of $\frac{1}{2}W_1$ distances to the barycenter gives the minimum classification error. Similarly, the W_2^2 distance under the ℓ_2 metric in Eq. (1) arises from the choice of the MSE loss. Second, our Eq. (3) recovers Eq. (2) in the noiseless setting, under which $p_a^* = r_a^*$ and $D_{\mathrm{TV}} = \frac{1}{2}W_1$. Denis et al. (2022) and Gaucher et al. (2022) also derived expressions for the tradeoff that resemble ours, but they assumed binary group or class labels.

Post-Processing. Given a biased model (e.g., pre-trained on biased data without constraints), this family of algorithms post-process the model to satisfy fairness, e.g., via remapping the outputs (Hardt et al., 2016; Pleiss et al., 2017). For DP fairness, existing procedures include but not limited to Fish et al. (2016); Menon and Williamson (2018); Chzhen et al. (2019); Jiang et al. (2020); Zeng et al. (2022a); Denis et al. (2022), but a major limitation is that they only handle binary group or class labels. In contrast, our proposed DP post-processing procedure works in the most general classification setting, and is accompanied by suboptimality and finite sample analyses. Moreover, it yields the optimal fair classifier when applied to the Bayes optimal predictor.

2 Preliminaries

Notation. Denote the $(k-1)$-dimensional probability simplex by $\Delta_k \coloneqq \{x \in \mathbb{R}^k : x \geq 0, \sum_{i=1}^k x_i = 1\}$, whose k vertices are $\{e_1, \cdots, e_k\}$, where $e_i \in \mathbb{R}^k$ is a vector of zeros and a single 1 on the i-th coordinate. Let $\mathcal{Q}_k$ denote the collection of distributions supported on the vertices of Δ_k. We will make heavy use of *randomized functions*, whose outputs follow a distribution conditioned on the input; a formal definition via the *Markov kernel* is provided in Definition B.2. Given a (randomized) function $f : \mathcal{X} \to \mathcal{Y}$ and a measure p on $\mathcal{X}$, we denote the *push-forward* of p by $f\sharp p$ (Definition B.3).

Problem Setup. Define a k-class classification problem by a joint distribution μ of input X, demographic group membership A (a.k.a. sensitive attribute), and class label Y, supported on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$; the labels may contain noise originated from e.g. data collection (Corbett-Davies and Goel, 2018). We assume a finite number of groups, $\mathcal{A} = [m] \coloneqq \{1, \cdots, m\}$, and use one-hot

representation for the class labels, i.e., $\mathcal{Y} = \{e_1, \cdots, e_k\}$. We denote the distribution under μ conditioned on group $A = a$ by μ_a, and the marginal distribution of input X by μ^X.

The *Bayes optimal predictor* on group a of the problem μ, denoted by $f^*_{a,\mu} : \mathcal{X} \to \Delta_k$, is a function that outputs the true class probabilities given each input x and $A = a$:

$$f^*_{a,\mu}(x)_i \coloneqq \mathbb{P}_{\mu_a}(Y = e_i \mid X = x) = \mathbb{E}_{\mu_a}[Y_i \mid X = x];$$

it coincides with the minimum MSE estimator of the one-hot labels Y given $(X, A = a)$. In addition, define $r^*_{a,\mu} \coloneqq f^*_{a,\mu}\sharp\mu^X_a$ to be the distribution of class probabilities on group a, supported on Δ_k; this quantity will appear throughout our presentation.

Fair classification aims to find a *randomized* and *attribute-aware* classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ on μ that achieves the minimum group-balanced classification error rate,[3]

$$\mathrm{err}(h; \mu) \coloneqq \frac{1}{m} \sum_{a \in [m]} \mathbb{P}(h(X, A) \neq Y \mid A = a) = \frac{1}{m} \sum_{a \in [m]} \int_{\mathcal{X} \times \mathcal{Y}} \mathbb{P}(h(x, a) \neq y) \, \mathrm{d}\mu_a(x, y), \tag{4}$$

under the constraints set by the designated fairness criteria; the decomposition on the r.h.s. is due to randomization.

We consider the group fairness criterion of demographic parity with relaxations:

Definition 2.1 (Demographic Parity with Relaxation)**.** Let $\epsilon \in [0, 1]$. A classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ satisfies ϵ-DP if $\Delta_{\mathrm{DP}}(h; \mu) \leq \epsilon$, where

$$\begin{aligned}\Delta_{\mathrm{DP}}(h; \mu) &\coloneqq \max_{a, a' \in [m]} \frac{1}{2} \sum_{y \in \mathcal{Y}} \left| \mathbb{P}(h(X, A) = y \mid A = a) - \mathbb{P}(h(X, A) = y \mid A = a') \right| \\ &= \max_{a, a' \in [m]} D_{\mathrm{TV}}(h\sharp(\mu^X_a \times \{a\}), h\sharp(\mu^X_{a'} \times \{a'\})),\end{aligned}$$

where $D_{\mathrm{TV}}(p, q) \coloneqq \frac{1}{2} \int |p(z) - q(z)| \, \mathrm{d}z$ denotes the total variation distance between measures, and

$$\mathbb{P}(h(X, A) = y \mid A = a) = \int_{\mathcal{X}} \mathbb{P}(h(x, a) = y) \, \mathrm{d}\mu^X_a(x)$$

is the proportion of outputs with label y on group a.

We call a classifier ϵ-*fair* when it satisfies ϵ-DP. The parameter ϵ controls the tradeoff between DP fairness and the maximal attainable accuracy; setting $\epsilon = 0$ recovers the standard strict definition of DP.[4]

Optimal Transport and Wasserstein Distance. Our study of fair classification involves optimal transports and Wasserstein distance, briefly reviewed below (Villani, 2003).

Definition 2.2 (Coupling)**.** Let p, q be probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. A coupling γ of p and q is a joint distribution on $\mathcal{X} \times \mathcal{Y}$ s.t. $p(x) = \int_{y \in \mathcal{Y}} \mathrm{d}\gamma(x, y)$ for all $x \in \mathcal{X}$, and $q(y) = \int_{x \in \mathcal{X}} \mathrm{d}\gamma(x, y)$ for all $y \in \mathcal{Y}$. Denote the collection of all couplings of p and q by $\Gamma(p, q)$.

[3] We extend our results to weighted errors in Appendix A.1, and consider balanced error in the main sections for clarity.

[4] The dependencies of f^*, r, err and Δ_{DP} on μ will be omitted.

Definition 2.3 (Optimal Transport)**.** Let p, q be probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ a cost function. The (Kantorovich) optimal transport from p to q under c, denoted by $\mathcal{T}^*_{p\to q,c} : \mathcal{X} \to \mathcal{Y}$, is a (randomized) function s.t. $\gamma^* \coloneqq (\mathrm{Id} \times \mathcal{T}^*_{p\to q,c})\sharp p \in \Gamma(p, q)$, where Id denotes the identity map, and achieves the optimal transportation cost, $\gamma^* \in \arg\inf_{\gamma\in\Gamma(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y)\,\mathrm{d}\gamma(x, y)$.

By definition, given $\gamma^* \in \Gamma(p, q)$ attaining the infimum above, we can derive an optimal transport, given by the randomized function satisfying $\mathbb{P}(\mathcal{T}^*_{p\to q,c}(X) = y \mid X = x) = \gamma^*(x, y)/\gamma^*(x, \mathcal{Y})$, $\forall x, y$.[5] Hence, the optimal transport is equivalently represented by the optimal coupling.

Lastly, when $\mathcal{X} = \mathcal{Y}$ and is a metric space equipped with distance d, the optimal transportation cost between p and q under $c = d$ coincides with their Wasserstein-1 distance:

Definition 2.4 (Wasserstein Distance)**.** Let p, q be probability measures on a metric space $(\mathcal{X}, d)$, and $r \in [1, \infty]$. The Wasserstein-r distance between p and q is $W_r(p, q) = (\inf_{\gamma\in\Gamma(p,q)} \int_{\mathcal{X}\times\mathcal{X}} d(x, x')^r\,\mathrm{d}\gamma(x, x'))^{1/r}$.

3 Fair and Optimal Classification

The first goal of this section is to give a characterization of the inherent tradeoff of DP fairness (Section 3.1), based on which, we will then build a DP post-processing procedure with suboptimality analyses (Sections 3.2 and 3.3).

3.1 Characterizing the Inherent Tradeoff

Our characterization is due to a reformulation of the classification problem assuming access to the Bayes optimal predictor. For any generic classification problem, we have that:

Lemma 3.1. *Let μ with $|\mathcal{A}| = 1$ be given along with the Bayes optimal predictor f^*, define $r^* \coloneqq f^*\sharp\mu^X$, and fix $q \in \mathcal{Q}_k$. Then for any (randomized) classifier $h : \mathcal{X} \to \mathcal{Y}$ satisfying $h\sharp\mu^X = q$, there exists a coupling $\gamma \in \Gamma(r^*, q)$ s.t.*

$$\mathrm{err}(h) = \frac{1}{2}\int_{\Delta_k\times\mathcal{Y}} \|s - y\|_1\,\mathrm{d}\gamma(s, y). \tag{5}$$

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

Under this reformulation, the problem of learning classifiers becomes equivalent to that of finding couplings γ and target output distributions q. Moreover, by exposing q, we can specify the output distribution to be satisfied by the classifier, under which, by Definition 2.4, the minimum error is

$$\min_{h:h\sharp\mu^X=q} \mathrm{err}(h) = \frac{1}{2}\min_{\gamma\in\Gamma(r^*,q)} \int_{\Delta_k\times\mathcal{Y}} \|s - y\|_1\,\mathrm{d}\gamma(s, y) = \frac{1}{2}W_1(r^*, q).$$

This is particularly convenient to the analysis of DP fairness, which is just a constraint on the output distributions, $q_a \coloneqq h\sharp(\mu^X_a \times \{a\})$, $\forall a \in [m]$. By Definition 2.1,

$$\Delta_{\mathrm{DP}}(h) \le \epsilon \iff \max_{a,a'\in[m]} D_{\mathrm{TV}}(q_a, q_{a'}) \le \epsilon.$$

[5] We will only consider transportation under the of ℓ_1 cost of $(x, y) \mapsto \|x - y\|_1$, hence omit the dependency of $\mathcal{T}^*_{p\to q}$ on c.

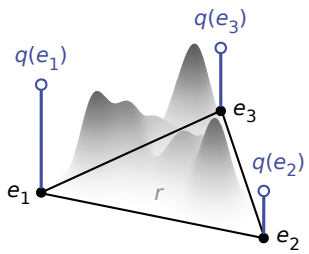

Figure 1: A distribution r on the 2d simplex Δ_3 (grey surface), and a distribution $q \in \mathcal{Q}_3$ on the vertices $\{e_1, e_2, e_3\}$ (blue spikes).

Therefore, for attribute-aware classifiers, whose components on each group, $h(\cdot, a)$, can be optimized independently, the results above directly give the following characterization of the minimum error rate under DP:

Theorem 3.2 (Minimum Error Rate Under DP)**.** *Let μ be given along with Bayes optimal predictor f_a^*'s, and $\epsilon \in [0, 1]$. Let $r_a^* \coloneqq f_a^* \sharp \mu_a^X$, $\forall a \in [m]$, then with W_1 under the ℓ_1 metric,*

$$\mathrm{err}_\epsilon^* \coloneqq \min_{h:\Delta_{\mathrm{DP}}(h)\le\epsilon} \mathrm{err}(h) = \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\le\epsilon}} \frac{1}{m}\sum_{a\in[m]} \frac{1}{2} W_1(r_a^*, q_a). \tag{6}$$

We describe it as a relaxed Wasserstein-barycenter problem of the r_a^*'s restricted to the vertices, because when $\epsilon = 0$, $q_1 = \cdots = q_m \in \mathcal{Q}_k$ and represent the barycenter under W_1 distance. It is a convex problem (in the primal form), and can be simplified given assumptions: in the noiseless setting with $\epsilon = 0$, it reduces to the TV-barycenter problem in Eq. (2). Zhao and Gordon (2022) established this equality for $m = k = 2$, whereas our reduction in Theorem A.2 holds generally.

Under strict DP ($\epsilon = 0$), the inherent tradeoff, namely the excess risk induced by the DP constraint, is

$$\frac{1}{2m}\left(\min_{q\in\mathcal{Q}_k} \sum_{a\in[m]} W_1(r_a^*, q) - \sum_{a\in[m]} \min_{q_a\in\mathcal{Q}_k} W_1(r_a^*, q_a)\right) \ge 0,$$

where the second term is the Bayes error rate and is achieved by the classifier $(x, a) \mapsto e_{\arg\max_i f_a^*(x)_i}$. So we may expect the tradeoff to be larger on problem instances where the r_a^*'s differ, and equal to zero when they are the same, or $\mathbb{E}_\mu[Y \mid X, A] \perp\!\!\!\perp A$, equivalently, since all groups would share the same optimal decision rule.

However, we point out that the tradeoff could be zero even if $\mathbb{E}_\mu[Y \mid X, A] \not\perp\!\!\!\perp A$, meaning that on such instances, enforcing DP would not degrade model performance. These scenarios arise from the nonuniqueness of the optimal classifier, which we illustrate in Example A.5.

Lastly, Zhao and Gordon (2022) concluded that in the noiseless setting, the tradeoff is zero iff the class marginals are the same, $\mathbb{E}_\mu[Y \mid A] \perp\!\!\!\perp A$. But this is no longer sufficient in the general noisy setting, by a counterexample (Example A.6).

3.2 Optimal Fair Classifier via Post-Processing

Theorem 3.2 only reformulates the fair classification problem, but the construction used in the proof of Lemma 3.1 for the equivalence (deferred to Appendix B) provides that the optimal fair classifier can be obtained by composing the Bayes optimal predictors f_a^* with post-processing functions.

More precisely, the post-processing functions are the optimal transport from r_a^*'s to the minimizing q_a^*'s of Eq. (6) (see Fig. 1 for a picture of these distributions):

Algorithm 1 Post-Process Predictor for ϵ-DP

Input: marginal input distributions $\mu_1^X, \cdots, \mu_m^X$, predictors $f_1, \cdots, f_m : \mathcal{X} \to \Delta_k$, relaxation $\epsilon \in [0, 1]$
$r_a \coloneqq f_a \sharp \mu_a^X$, $\forall a \in [m]$
$q_1, \cdots, q_m \leftarrow$ minimizer of Eq. (6) on $r_1, \cdots, r_m, \epsilon$ ▷ *relaxed barycenter problem*
for $a = 1$ **to** m **do**
    $\mathcal{T}^*_{r_a \to q_a} \leftarrow r_a$ to q_a optimal transport under ℓ_1 cost
end for
Return: $(x, a) \mapsto \mathcal{T}^*_{r_a \to q_a} \circ f_a(x)$

Theorem 3.3 (Optimal Classifier Under DP)**.** *Let $\epsilon \in [0, 1]$, $q_1^*, \cdots, q_m^*$ be a minimizer of Eq.* (6)*, and $\mathcal{T}^*_{r_a^* \to q_a^*}$ the optimal transport from r_a^* to q_a^* under the ℓ_1 cost. Then*

$$(x, a) \mapsto \mathcal{T}^*_{r_a^* \to q_a^*} \circ f_a^*(x) \in \operatorname*{arg\,min}_{h : \Delta_{\mathrm{DP}}(h) \le \epsilon} \operatorname{err}(h).$$

To summarize, optimal fair classifiers can be computed in 3 steps: (1) learn f_a^* by minimizing MSE w.r.t. the one-hot Y,

$$f_a^* = \operatorname*{arg\,min}_{f : \mathcal{X} \to \Delta_k} \mathbb{E}_\mu \big[\|f(X) - Y\|_2^2 \mid A = a \big], \tag{7}$$

(2) find the minimizing q_a^*'s of Eq. (6), the barycenter problem, and (3) compute the optimal transports $\mathcal{T}^*_{r_a^* \to q_a^*}$, then return $(x, a) \mapsto \mathcal{T}^*_{r_a^* \to q_a^*} \circ f_a^*(x)$. The last two steps constitute the post-processing, reproduced in Algorithm 1.

In practice, however, the f_a^*'s may not be exactly learned due to cost, or difficulties in representation, optimization, and generalization. Instead, we will often work with suboptimal predictors $f_a \approx f_a^*$ (pre-trained by a vendor). The following section shows that in these cases, Algorithm 1 is still applicable to post-processing the f_a's for fairness.

3.3 Post-Processing Pre-Trained Predictors

Let $f_1, \cdots, f_m : \mathcal{X} \to \Delta_k$ be arbitrary (randomized) predictors on each group. We want to find post-processing functions $g_a : \Delta_k \to \mathcal{Y}$ such that when applied to f_a, the derived classifier $(x, a) \mapsto g_a \circ f_a(x)$ satisfies fairness, and ideally achieves the minimum error rate among all fair classifiers derived from the f_a's.

The classifier $\bar{h}(x, a) \coloneqq \mathcal{T}^*_{r_a \to q_a} \circ f_a(x)$ (where $r_a \coloneqq f_a \sharp \mu_a^X$) returned from calling Algorithm 1 on the f_a's is ϵ-fair, because by construction, $\bar{h} \sharp (\mu_a^X \times \{a\}) = q_a$ for all a, and $\Delta_{\mathrm{DP}}(\bar{h}) = \max_{a,a'} D_{\mathrm{TV}}(q_a, q_{a'}) \le \epsilon$ from the constraint in Eq. (6). The optimality depends on the L^1 error of the f_a's w.r.t. the Bayes optimal predictors f_a^*:

Theorem 3.4 (Error Propagation)**.** *Let predictor f_a's be given, $\epsilon \in [0, 1]$, and $\operatorname{err}_\epsilon^*$ as in Eq.* (6)*. For the ϵ-fair classifier $\bar{h}$ returned from Algorithm 1,*

$$0 \le \operatorname{err}(\bar{h}) - \operatorname{err}_\epsilon^* \le \frac{1}{m} \sum_{a \in [m]} \mathbb{E}_{\mu_a^X} [\|f_a(X) - f_a^*(X)\|_1].$$

Hence, while the decrease in performance of the post-processed $\bar{h}$ when $f_a = f_a^*$ the Bayes optimal predictor is attributed entirely to the inherent tradeoff (Theorem 3.3), this may not be

the case if $f_a \neq f_a^*$, due to information loss. But, to guarantee that the returned $\bar{h}$ is (at least) optimal among all fair classifiers derived from f_a, we require them to be *calibrated* prior to calling Algorithm 1:

Definition 3.5 (Calibration)**.** Let μ with $|\mathcal{A}| = 1$ be given. The predictor $f : \mathcal{X} \to \Delta_k$ is said to be calibrated if $\mathbb{P}_\mu(Y = e_i \mid f(X) = s) = s_i$ for all $s \in \Delta_k$ and $i \in [k]$.

The predictors f_a's can be calibrated by composing with calibration maps $u_a : \Delta_k \to \Delta_k$ as extra post-processing. The optimal calibration maps u_a^*'s (in the sense that no further information loss is incurred) are by definition the minimum MSE estimators of Y given $(f_a(X), A = a)$, which can be learned from labeled data. When the f_a's are already calibrated (e.g., Bayes optimal predictors), $u_a^* = \mathrm{Id}$ identity map. Alternatively, calibration could be achieved with more efficient methods, e.g., Guo et al. (2017) calibrates neural networks via histogram binning.

Note that finding the optimal fair post-processing functions on the f_a's is equivalent to finding the optimal fair classifier on a new problem μ' derived from the original μ under an input transformation—the joint distribution of $(f_A(X), A, Y)$. Also, the Bayes optimal predictors on μ' coincide with the optimal calibration maps, $\mathbb{E}_{\mu'}[Y \mid X' = s, A = a] = u_a^*(s)$. Hence, by Theorem 3.3, the optimal fair classifier on μ' equal the post-processing maps found by Algorithm 1 on the optimally calibrated predictors, $u_a^* \circ f_a$, which implies that the returned $\bar{h}$ is optimal among all derived fair classifiers. Lastly, we remark that the suboptimality due to miscalibration can also be bounded using Theorem 3.4 after replacing f_a^* in the statement with $u_a^* \circ f_a$.

4 Algorithms for Finite Samples

In this section, we instantiate our proposed DP post-processing procedure, Algorithm 1, to scenarios where only finite (unlabeled) samples are available:

Assumption 4.1. We have i.i.d. samples of (X, A) in the form of $S_a \coloneqq (x_{a,i})_{i=1}^n$, $x_{a,i} \sim \mu_a^X$, $\forall a \in [m]$, which are independent of the predictor f_a's being post-processed.

We provide finite sample guarantees for both the fairness and the error rate of the returned classifier; the latter is stated in terms of the suboptimality relative to the optimal fair classifier derived from calibrated versions of the predictors:

Assumption 4.2. The predictor f_a's being post-processed are *calibrated*. Let $r_a \coloneqq f_a \sharp \mu_a^X$, by discussions in Section 3.3, the optimal derived ϵ-fair classifier has error rate

$$\mathrm{err}_{\epsilon,f}^* \coloneqq \min_{\substack{q_1,\cdots,q_m \in \mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a, q_{a'}) \leq \epsilon}} \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} W_1(r_a, q_a).$$

4.1 The Finite Case

We begin with the case where the $\mathrm{supp}(r_a) =: \mathcal{R}_a$'s are finite (but the input distribution μ^X need not be finite). Note that it covers the scenario of post-processing pre-trained *classifiers*, $f_a : \mathcal{X} \to \mathcal{Y}$, which have a finite output space.

If the true probability mass of the r_a's are known, then Algorithm 1 is instantiated by a linear

program (LP):

$$
\begin{aligned}
\mathrm{OPT}(r_1, \cdots, r_m, \epsilon): \quad \min_{\substack{q_1,\cdots,q_m \\ \gamma_1,\cdots,\gamma_m}} \quad & \sum_{a\in[m]} \sum_{s\in\mathcal{R}_a, y\in\mathcal{Y}} \|s-y\|_1 \cdot \gamma_a(s,y) & \\
\text{s.t.} \quad & \sum_{y'\in\mathcal{Y}} \left|q_a(y') - q_{a'}(y')\right| \le \epsilon, & \forall a, a' \in [m], \\
& \sum_{y'\in\mathcal{Y}} \gamma_a(s,y') = r_a(s), & \forall a \in [m],\ s \in \mathcal{R}_a, \\
& \sum_{s'\in\mathcal{R}_a} \gamma_a(s',y) = q_a(y), & \forall a \in [m],\ y \in \mathcal{Y}, \\
& \sum_{y'\in\mathcal{Y}} q_a(y') = 1, \ \ q_a \ge 0, \ \ \gamma_a \ge 0, & \forall a \in [m].
\end{aligned}
$$

where $q_1, \cdots, q_m \in \mathbb{R}^k$, and $\gamma_a \in \mathbb{R}^{|\mathcal{R}_a|\times k}$. This program simultaneously finds the minimizers $q_1^*, \cdots, q_m^*$ of the relaxed barycenter problem in Eq. (6) and the optimal transports in Theorem 3.3, which are stored in the solution $\gamma_1^*, \cdots, \gamma_m^*$ of OPT: each $\mathcal{T}^*_{r_a\to q^*}$ is a randomized function satisfying $\mathbb{P}(\mathcal{T}^*_{r_a\to q^*}(R) = y \mid R = s) = \gamma_a^*(s,y) / \sum_{y'\in\mathcal{Y}} \gamma_a^*(s,y')$, for all $s \in \mathcal{R}_a$.

If the pmfs of the r_a's are unknown but finite samples S_a are given, then we proceed with the empirical pmfs, $\hat{r}_a \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{f_a(x_{a,i})}$, where δ denotes the Dirac delta function, by solving $\mathrm{OPT}(\hat{r}_1, \cdots, \hat{r}_m, \epsilon)$ to obtain estimated $\hat{q}_1, \cdots, \hat{q}_m$, and use the empirical transports $\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}$ to post-process the classifier, $\hat{h}(x,a) = \mathcal{T}^*_{\hat{r}_a\to\hat{q}_a} \circ f_a(x)$ (assign arbitrary label to unseen inputs during inference, $f_a(x) \notin f_a(S_a)$). It has the following guarantees:

Theorem 4.3 (Generalization, Finite Case)**.** *Let predictor f_a's be given, and $\epsilon \in [0,1]$. Under Assumption 4.1, let $|\mathcal{R}| \coloneqq \max_a |\mathrm{supp}(r_a)|$, w.p. at least $1-\delta$ over the random draw of the samples, for the classifier $\hat{h}$ derived from above,*

$$
\Delta_{\mathrm{DP}}(\hat{h}) \le \epsilon + \sqrt{\frac{32|\mathcal{R}|}{n}\left(\frac{1}{4} + \ln\left(\frac{2m}{\delta}\right)\right)},
$$

and additionally with Assumption 4.2,

$$
\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\epsilon,f} \le \sqrt{\frac{72|\mathcal{R}|}{n}\left(\frac{1}{4} + \ln\left(\frac{2m}{\delta}\right)\right)}.
$$

4.2 The Continuous Case

When the r_a's are continuous,[6] we may still use the program $\mathrm{OPT}(\hat{r}_1, \cdots, \hat{r}_m, \epsilon)$ to estimate the optimal target output distributions, $\hat{q}_1, \cdots, \hat{q}_m$, but the empirical transports $\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}$ are no longer usable for post-processing here, since the inputs at inference time will be unseen almost surely ($f_a(x) \notin f_a(S_a)$). Instead, after obtaining the $\hat{q}_a$'s, we need to approximate the optimal transports $\mathcal{T}^*_{r_a\to\hat{q}_a}$ from r_a's (population) to the $\hat{q}_a$'s from finite samples for all $a \in [m]$.

Note that this is a semi-discrete optimal transport problem, studied in a long line of work (Genevay et al., 2016; Staib et al., 2017; Chen et al., 2019), where a common procedure is to reformulate

[6] I.e., the probability measure does not give mass to sets whose intersection with Δ_k has Hausdorff dimension less than $k-1$.

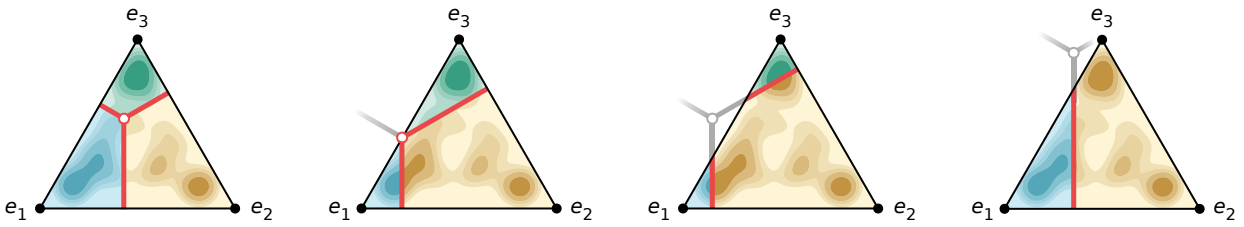

Figure 2: Examples of simplex-vertex optimal transports for $k = 3$ (vertex distributions are different). All points in the lower-left blue partition are transported to e_1, lower-right yellow to e_2, and upper green to e_3. The transports are described by a Y-shaped boundary.

optimal transport as a convex optimization problem over a vector $\psi_a \in \mathbb{R}^k$ using the Kantorovich-Rubinstein dual and the c-transform of the Kantorovich potential ϕ_a. Namely, for each $a \in [m]$,

$$\begin{aligned}\inf_{\gamma_a \in \Gamma(r_a, \hat{q}_a)} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma_a(s, y) &= \sup_{\substack{\phi_a : \Delta_k \to \mathbb{R},\, \psi_a \in \mathbb{R}^k \\ \phi_a(s) + \psi_{a,i} \le \|s - e_i\|_1}} \left(\int_{\Delta_k} \phi_a(s) \, \mathrm{d}r_a(s) + \sum_{i=1}^{k} \psi_{a,i} \hat{q}_a(e_i) \right) \\ &= \sup_{\psi_a \in \mathbb{R}^k} \left(\int_{\Delta_k} \min_i (\|s - e_i\|_1 - \psi_{a,i}) \, \mathrm{d}r_a(s) + \sum_{i=1}^{k} \psi_{a,i} \hat{q}_a(e_i) \right),\end{aligned}$$

whereby the optimal ψ_a^* can be found by e.g. (stochastic) gradient ascent w.r.t. the last expression. Moreover, Gangbo and McCann (1996) showed that in the semi-discrete case, the optimal transport $\mathcal{T}^*_{r_a \to \hat{q}_a}$ belongs to the class of deterministic functions of

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

(break ties to the tied e_i with the largest index i); specifically, it is equal to the function with parameter ψ_a^*. Pictures of semi-discrete optimal transports for $k = 3$ are included in Fig. 2.

The proof of Gangbo and McCann (1996) required the cost function to be strictly convex and superlinear (Assumptions H1 to H3 in their paper), which are not satisfied by our ℓ_1 cost (Ambrosio and Pratelli, 2003). Therefore, for our simplex-vertex transportation problem under the ℓ_1 cost, we give a proof in Appendix D for the existence and uniqueness of the optimal transport in $\mathcal{G}_k$ via analyzing its geometry.

Our Implementation. The optimal transports $\mathcal{T}^*_{r_a \to \hat{q}_a}$ can be estimated via solving an extra set of optimization problems subsequent to obtaining the $\hat{q}_a$'s from OPT. But curiously, is it possible to eliminate this step, and instead, extract a set of transport mappings directly from the solution γ_a^* to OPT (equivalently, $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$) good enough for post-processing? Yes!

Algorithm 2 details our post-processing procedure for finite samples in the continuous case. We highlight Lines 6–9, where transport mappings $\mathcal{T}_a$ are extracted from empirical transports $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$ returned from OPT; a step-by-step illustration is provided in Fig. 3, and formal derivations are deferred to Appendix D, which arise from the geometric analysis of the optimal transport. Each $\mathcal{T}_a$ has the property that it agrees with $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$ on all points in S_a that do not lie on the boundaries of $\mathcal{T}_a$. Since the boundaries are described by at most $\binom{k}{2}$ hyperplanes, the number of disagreements is almost surely $O(k^2)$. And thanks to the low complexity of $\mathcal{G}_k \ni \mathcal{T}_a$ (Theorem D.2), classifiers derived using Algorithm 2 have the following guarantees:

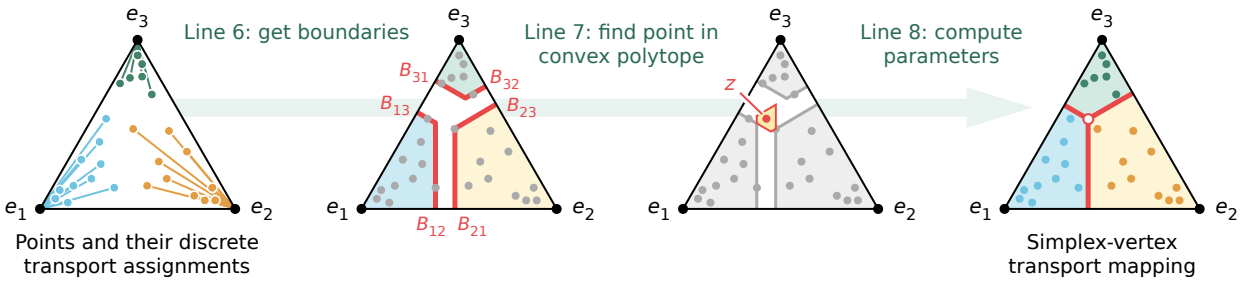

Figure 3: Extract a simplex-vertex transport $\mathcal{T} \in \mathcal{G}_3$ that agrees with the discrete optimal transport (except for points that lie on the boundaries), corresponding to Lines 6–9 of Algorithm 2. For illustration, the discrete transport on the left does not split mass.

Algorithm 2 Post-Process Predictor for ϵ-DP (Finite Samples, Continuous Case)

Input: unlabeled samples $S_1, \cdots, S_m$, predictors $f_1, \cdots, f_m : \mathcal{X} \to \Delta_k$, relaxation $\epsilon \in [0, 1]$
$\hat{r}_a \coloneqq \frac{1}{|S_a|} \sum_{x \in S_a} \delta_{f_a(x)}$, $\forall a \in [m]$
$\gamma_1, \cdots, \gamma_m \leftarrow$ minimizer of $\mathrm{OPT}(\hat{r}_1, \cdots, \hat{r}_m, \epsilon)$ ▷ *relaxed barycenter problem in finite case*
$v_{ij} \coloneqq e_j - e_i$
for $a = 1$ **to** m **do** ▷ *extract transports*
  $B_{ij} \leftarrow \max\{f_a(x)^\top v_{ij} + 1 : x \in S_a, \gamma_a(f_a(x), e_i) > 0\} \cup \{0\}$
  $z \leftarrow$ point in $\bigcap_{i \neq j} \{x \in \mathbb{R}^k : x^\top v_{ij} \geq B_{ij} - 1\}$ ▷ *arbitrary; always exists*
  $\psi_i \leftarrow 2 z^\top v_{i1}$, $\forall i \in [k]$
  $\mathcal{T}_a \leftarrow (s \mapsto e_{\arg\min_{i \in [k]} (\|s - e_i\|_1 - \psi_i)})$
end for
Return: $(x, a) \mapsto \mathcal{T}_a \circ f_a(x)$

Theorem 4.4 (Generalization, Continuous Case)**.** *Let predictor f_a's be given, and $\epsilon \in [0, 1]$. Under Assumption 4.1, w.p. at least $1 - \delta$ over the random draw of the samples, for the classifier $\hat{h}$ returned from Algorithm 2,*

$$\Delta_{\mathrm{DP}}(\hat{h}) \leq \epsilon + O\left(\sqrt{\frac{k^3}{n} \ln\left(\frac{mk}{\delta}\right)} + \frac{k^2}{n}\right),$$

and additionally with Assumption 4.2,

$$\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\epsilon,f} \leq O\left(\sqrt{\frac{k}{n} \ln\left(\frac{mn}{k\delta}\right)} + \frac{k^2}{n}\right).$$

The first term in both expressions is from standard sample complexity analysis, and the second term is attributed to the aforementioned disagreements.

4.3 The General Case

For completeness, we briefly discuss post-processing in the general case where the r_a's are neither finite nor purely continuous (i.e., they contain atoms) via distribution smoothing.

Let ρ_s be a continuous distribution[6] with finite first moment (potentially depends on s), and u_ρ a randomized function s.t. $u_\rho(s) \sim s + N$ with $N \sim \rho_s$ as i.i.d. random noise. Let q_a's denote the minimizer of the relaxed barycenter problem in Eq. (6) on the r_a's with ϵ, and $\mathcal{T}^*_{\tilde{r}_a \to q_a}$ the optimal

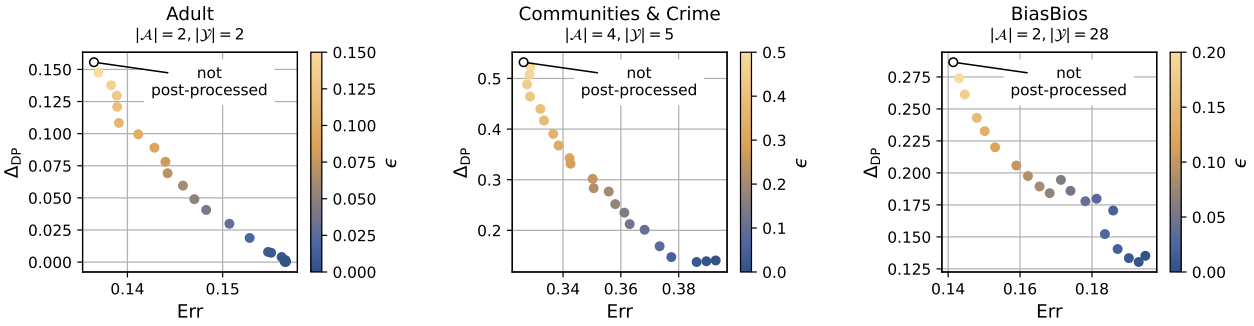

Figure 4: Group-balanced error rate (Eq. (4)) and DP fairness (Δ_{DP}; Definition 2.1) of classifiers post-processed using Algorithm 2, under varying settings for ϵ. The number of groups $\mathcal{A}$ and classes $\mathcal{Y}$ on each task are included.

transport from $\tilde{r}_a := u_\rho \sharp r_a$ to q_a (can be obtained with the method in Algorithm 2 if $\mathrm{supp}(\tilde{r}_a) \subseteq \Delta_k$). Then the classifier given by $\bar{h}_\rho(x, a) := \mathcal{T}^*_{\tilde{r}_a \to q_a} \circ u_\rho \circ f_a(x)$ is ϵ-fair, and its suboptimality is controlled via the *bandwidth* of ρ_s:

Theorem 4.5 (Error Propagation, Smoothing)**.** *Let predictor f_a's be given, and $\epsilon \in [0, 1]$. Under Assumption 4.2, for the ϵ-fair classifier $\bar{h}_\rho$ derived from above,*

$$0 \le \mathrm{err}(\bar{h}_\rho) - \mathrm{err}^*_{\epsilon,f} \le \frac{1}{m} \sum_{a \in [m]} \mathbb{E}_{N \sim \rho_s, s \sim r_a}[\|N\|_1].$$

E.g., if $\rho = \mathrm{Laplace}(0, b \cdot I_k)$, then $\mathbb{E}[\|N\|_1] = kb$, and the suboptimality attributed to smoothing is less than kb.

5 Experiments

To verify and demonstrate the effectiveness of our post-processing procedure (Algorithm 1), we apply Algorithm 2 (compute OPT with LP solvers) on three real-world benchmark datasets: the Adult (Kohavi, 1996) and Communities & Crime tabular datasets (Redmond and Baveja, 2002), and the BiasBios text dataset (De-Arteaga et al., 2019). The pre-trained attribute-aware predictors on which Algorithm 2 is applied are trained to minimize MSE of the one-hot class labels *without constraints*. We use linear predictors (learned via OLS) on the tabular datasets, and a fine-tuned BERT base neural language model on BiasBios (Devlin et al., 2019). The training data are split for pre-training and post-processing. Descriptions of the datasets, the classification tasks, and further experiment details are in Appendix E.

Results. Figure 4 presents the test set evaluation results of post-processed classifiers returned from Algorithm 2 under varying settings of ϵ. On all datasets, the post-processing procedure reduces the bias (i.e. Δ_{DP}) without significant increases to the error rate, and their tradeoff is effectively adjusted via ϵ. Note that almost precise control of Δ_{DP} is achieved on the Adult dataset, which contains sufficient number of (unlabeled) training data for generalization.

Under strict DP ($\epsilon = 0$), near-zero Δ_{DP} is attained on Adult (see Table 2 for numbers), but not on the other two datasets due to generalization error (Communities & Crime is a small dataset, and BiasBios has many classes), or potentially the violation of the continuous assumption required by Algorithm 2 (Section 4.2). The latter could be remedied by, e.g., the smoothing procedure

discussed in Section 4.3. We perform smoothing using a Dirichlet distribution, and observe further improvements to Δ_{DP} (results deferred to Table 2 of Appendix E.3). Lastly, while not pursued here, lower error rates are possible if the predictors are calibrated prior to post-processing, as discussed in Section 3.3.

6 Further Related Work

Fairness criteria are generally categorized into individual fairness (Dwork et al., 2012; Sharifi-Malvajerdi et al., 2019), subgroup (Kearns et al., 2018), and group fairness (Zemel et al., 2013; Hardt et al., 2016; Kleinberg et al., 2017; Verma and Rubin, 2018). A predictive model is said to satisfy individual fairness if it treats *similar* individuals similarly; the practical challenge with this notion of fairness lies in the specification of the similarity measure, which needs to be application and context dependent. On the other hand, (sub)group fairness, to which demographic parity belongs, is defined on population-level statistics, e.g., true positive and/or negative rates (Hardt et al., 2016), predictive rates (Chouldechova, 2017; Berk et al., 2021; Zeng et al., 2022b), accuracy (Buolamwini and Gebru, 2018), etc.

For mitigating the bias in machine learning models under the fairness criteria, a variety of algorithms have been proposed. Besides post-processing, there are methods based on data pre-processing (Calmon et al., 2017; Song et al., 2019), constrained optimization (Kamishima et al., 2012; Zafar et al., 2017; Agarwal et al., 2019), and fair representation learning (Zemel et al., 2013; Madras et al., 2018; Zhao et al., 2020). There are also methods under other learning paradigms, such as unsupervised learning (Chierichetti et al., 2017; Backurs et al., 2019; Li et al., 2020), ranking (Zehlike et al., 2017), and sequential decision making (Joseph et al., 2016, 2017; Gillen et al., 2018; Chi et al., 2022).

7 Conclusion

In this paper, we characterized the inherent tradeoff of DP fairness on classification problems in the most general setting, and proposed a post-processing procedure with generalization guarantees. Our implementation uses LP solvers; while they enjoy stability and consistent performance, a potential concern is scalability to larger numbers of classes or samples. It would be of practical value to analyze an implementation that uses more runtime-efficient optimization methods, e.g., gradient descent.

Technically, we studied the geometry of the optimal transport between distributions supported on the simplex and its vertices. A main result is that when the distributions are semi-discrete, the optimal transport is unique, and is given by the c-transform of the Kantorovich potential, including under the ℓ_1 cost. This result may be of independent theoretical interest to the community.

Our results add to the line of work that study the inherent tradeoffs of fairness criteria, which we believe would benefit practitioners in the design of fair machine learning systems, and contribute to a better understanding of the implications of fairness in machine learning.

Acknowledgements

The authors thank Jane Du, Yuzheng Hu, and Seiyun Shin for feedback on the draft. 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*, volume 1813 of *Lecture Notes in Mathematics*, pages 123–160. Springer, 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: Limitations and Opportunities*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

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.AT]*.

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 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.LG]*.

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.

Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, 2017.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *Advances in Neural Information Processing Systems*, 2019.

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

Sam Corbett-Davies and Sharad Goel. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning, 2018. *arXiv:1808.00023 [cs.CY]*.

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.ST]*.

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.

Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A Confidence-Based Approach for Balancing Fairness and Accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*, pages 144–152, 2016.

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 [stat.ML]*.

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.

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.LG]*.

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.

Jonas Moritz Kohler and Aurelien Lucchi. Sub-sampled Cubic Regularization for Non-convex Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1895–1904, 2017.

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

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.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.

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.

Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel Streaming Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, 2017.

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.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

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, Canwen 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.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-Optimal Classifiers under Group Fairness, 2022a. *arXiv:2202.09724 [stat.ML]*.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Fair Bayes-Optimal Classifiers Under Predictive Parity. In *Advances in Neural Information Processing Systems*, 2022b.

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 on Theorem 3.2

This section contains deferred discussions on Theorem 3.2: the extension to arbitrarily group-weighted classification error (Appendix A.1), the reduction to TV-barycenter in the noiseless setting (Appendix A.2), and providing examples to the remarks (Appendix A.3).

A.1 Group-Weighted Classification Error Rates

For clarity, in the main sections, we focused on group-balanced classification error. But our results apply generally to arbitrarily group-weighted error via an extension of Theorem 3.2.

Let $w \in \mathbb{R}^m_{\geq 0}$ denote a set of nonnegative weights assigned to each group, and define group-weighted classification error under w by

$$\mathrm{err}_w(h) \coloneqq \sum_{a\in[m]} w_a \cdot \mathbb{P}(h(X,A) \neq Y \mid A = a).$$

Theorem A.1 (Full Version of Theorem 3.2)**.** *Let μ be given along with Bayes optimal predictor f_a^*'s, $\epsilon \in [0,1]$, and $w \in \mathbb{R}^m_{\geq 0}$. Let $r_a^* \coloneqq f_a^* \sharp \mu_a^X$, $\forall a \in [m]$, then with W_1 under the ℓ_1 metric,*

$$\min_{h:\Delta_{\mathrm{DP}}(h)\leq\epsilon} \mathrm{err}_w(h) = \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\leq\epsilon}} \sum_{a\in[m]} w_a \cdot \frac{1}{2} W_1(r_a^*, q_a).$$

Proof. Lemma 3.1 implies that for each $a \in [m]$ and fixed $q_a \in \mathcal{Q}_k$, the minimum error rate on group a, denoted by err_a, among randomized classifiers $h_a : \mathcal{X} \to \mathcal{Y}$ whose output distribution equals to q_a, is given by

$$\min_{h_a:h_a\sharp\mu_a^X=q_a} \mathrm{err}_a(h_a) \coloneqq \min_{h_a:h_a\sharp\mu_a^X=q_a} \mathbb{P}(h_a(X) \neq Y \mid A = a) = \frac{1}{2} W_1(r_a^*, q_a).$$

Because of attribute-awareness, we can optimize each component of $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$, $h(\cdot, a) =: h_a$ for all $a \in [m]$, independently. So for any set of fixed $q_1, \cdots, q_m \in \mathcal{Q}_k$,

$$\min_{h:h_a\sharp\mu_a^X=q_a,\forall a} \mathrm{err}_w(h) = \sum_{a\in[m]} w_a \cdot \min_{h_a:h_a\sharp\mu_a^X=q_a} \mathrm{err}_a(h_a) = \sum_{a\in[m]} w_a \cdot \frac{1}{2} W_1(r_a^*, q_a).$$

Incorporating the ϵ-DP constraint, we get

$$\begin{aligned}\min_{h:\Delta_{\mathrm{DP}}(h)\leq\epsilon} \mathrm{err}_w(h) &= \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\leq\epsilon}} \min_{h:h_a\sharp\mu_a^X=q_a,\forall a} \mathrm{err}_w(h) \\ &= \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\leq\epsilon}} \sum_{a\in[m]} w_a \cdot \frac{1}{2} W_1(r_a^*, q_a).\end{aligned}$$

□

A.2 Reduction to TV-Barycenter in Noiseless Setting

When the classification problem μ is noiseless, i.e., the (unconstrained) Bayes error rate is zero, $\min_h \mathrm{err}(h) = 0$, we show that Theorem 3.2 reduces to a relaxed TV-barycenter problem. For strict DP ($\epsilon = 0$), this result is previously established by Zhao and Gordon (2022) for the binary case of $m = k = 2$. Our reduction here holds for the general case.

Note that noiselessness means that there exist deterministic labeling functions $y_a : \mathcal{X} \to \mathcal{Y}$ for each $a \in [m]$ s.t. $Y = y_A(X)$ almost surely. Therefore, the Bayes optimal predictors $f_a^* = y_a$, and we have

$$r_a^* = p_a^* \coloneqq f_a^* \sharp \mu_a^X \quad \text{where} \quad p_a^*(e_i) = \mathbb{P}_\mu(Y = e_i \mid A = a).$$

Theorem A.2 (Minimum Error Rate Under DP, Noiseless Setting)**.** *Let noiseless μ be given, and $\epsilon \in [0, 1]$. Then*

$$\begin{aligned}\min_{h:\Delta_{\mathrm{DP}}(h)\le\epsilon} \mathrm{err}(h) &= \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\le\epsilon}} \frac{1}{m}\sum_{a\in[m]} \frac{1}{2} W_1(r_a^*, q_a) \\ &= \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a,q_{a'})\le\epsilon}} \frac{1}{m}\sum_{a\in[m]} D_{\mathrm{TV}}(p_a^*, q_a).\end{aligned}$$

This is due to $\mathrm{supp}(p_a^*) \subseteq \{e_1, \cdots, e_k\}$ sharing the same finite support with any $q_a \in \mathcal{Q}_k$, whereby $\frac{1}{2}W_1(r_a^*, q_a) = \frac{1}{2}W_1(p_a^*, q_a) = D_{\mathrm{TV}}(p_a^*, q_a)$. Specifically, recall the fact that W_1 under the 0-1 distance is equal to D_{TV}:

Proposition A.3. *Let p, q be probability measures on $\mathcal{X}$ with metric $d(x, y) = \mathbb{1}(x \neq y)$, where $\mathbb{1}(\cdot)$ denotes the indicator function. Then $W_1(p, q) = D_{\mathrm{TV}}(p, q)$.*

Proof. By definition,

$$\begin{aligned}W_1(p,q) &= \inf_{\gamma\in\Gamma(p,q)} \int_{\mathcal{X}\times\mathcal{X}} \mathbb{1}(x \neq y)\,\mathrm{d}\gamma(x,y) \\ &= \left(1 - \sup_{\gamma\in\Gamma(p,q)} \int_{\mathcal{X}\times\mathcal{X}} \mathbb{1}(x = y)\,\mathrm{d}\gamma(x,y)\right) \\ &= \left(1 - \int_{\mathcal{X}} \min(p(x), q(x))\,\mathrm{d}x\right) \\ &= \int_{\mathcal{X}} \max(0, q(x) - p(x))\,\mathrm{d}x \\ &= \frac{1}{2}\int_{\mathcal{X}} |q(x) - p(x)|\,\mathrm{d}x =: D_{\mathrm{TV}}(p,q),\end{aligned}$$

where line 3 is due to $\gamma(x, x) \le \min(p(x), q(x))$ for all $\gamma \in \Gamma(p, q)$ and there always exists a coupling s.t. $\gamma(x, x) = \min(p(x), q(x))$, and line 5 to $\int_{\mathcal{X}} q(x) - p(x)\,\mathrm{d}x = 0$. □

Proof of Theorem A.2. Because $\mathrm{supp}(r_a^*) = \mathrm{supp}(p_a^*) \subseteq \{e_1, \cdots, e_k\}$, the ℓ_1 distance (in W_1) between any $s \in \mathrm{supp}(r_a^*)$ and $y \in \{e_1, \cdots, e_k\}$ simplifies to $\|s - y\|_1 = 2 \cdot \mathbb{1}(s \neq y)$. The result then follows from Proposition A.3. □

Moreover, in this case, we have closed-form solution for the optimal fair classifier in Theorem 3.3:

Theorem A.4 (Optimal Classifier Under DP, Noiseless Setting)**.** *Let noiseless μ be given along with the ground-truth labeling functions y_a, and let $\epsilon \in [0, 1]$. Let $q_1^*, \cdots, q_m^*$ be a minimizer of Eq.* (6)*, define*

$$d_a(e_i) \coloneqq \frac{\max(0, q_a^*(e_i) - p_a^*(e_i))}{\sum_{j\in[k]} \max(0, q_a^*(e_j) - p_a^*(e_j))}, \quad s_a(e_i) \coloneqq \frac{\max(0, p_a^*(e_i) - q_a^*(e_i))}{p_a^*(e_i)}, \quad \forall a \in [m],\, i \in [k],$$

and randomized functions $\mathcal{T}_a : \mathcal{Y} \to \mathcal{Y}$ *for each* $a \in [m]$ *satisfying*

$$\mathcal{T}_a(e_i) = \begin{cases} e_i & w.p.\ s_a(e_i)d_a(e_i) + (1 - s_a(e_i)), \\ e_j & w.p.\ s_a(e_i)d_a(e_j), \qquad\qquad \forall j \in [m],\ j \neq i. \end{cases}$$

Then

$$(x, a) \mapsto \mathcal{T}_a \circ y_a(x) \in \operatorname*{arg\,min}_{h:\Delta_{\mathrm{DP}}(h)\leq\epsilon} \operatorname{err}(h).$$

Proof. We first verify that $(\mathcal{T}_a \circ y_a)\sharp\mu_a^X = q_a^*$. For all $e_i \in \mathcal{Y}$,

$$\begin{aligned} \mathbb{P}(\mathcal{T}_a \circ y_a(X) = e_i \mid A = a) &= p_a^*(e_i)(s_a(e_i)d_a(e_i) + (1 - s_a(e_i))) + d_a(e_i) \sum_{i\neq j} p_a^*(e_j)s_a(e_j) \\ &= p_a^*(e_i)(1 - s_a(e_i)) + d_a(e_i) \sum_{j\in[k]} p_a^*(e_j)s_a(e_j) \\ &= p_a^*(e_i) - \max(0, p_a^*(e_i) - q_a^*(e_i)) + \max(0, q_a^*(e_i) - p_a^*(e_i)) \\ &= q_a^*(e_i), \end{aligned}$$

where line 3 is because $\sum_{i\in[k]} p_a^*(e_i) - q_a^*(e_i) = 0 \implies \sum_{i\in[k]} \max(0, p_a^*(e_i) - q_a^*(e_i)) = \sum_{i\in[k]} \max(0, q_a^*(e_i) - p_a^*(e_i))$, and the last line is from case analysis.

Next, we compute the error rate on group $a \in [m]$. By construction, its accuracy conditioned on $Y = y_a(X) = e_i$ is

$$\mathbb{P}(\mathcal{T}_a \circ y_a(X) = e_i \mid A = a, Y = e_i) = \begin{cases} 1 - s_a(e_i) & \text{if } d_a(e_i) = 0, \\ 1 = 1 - s_a(e_i) & \text{if } d_a(e_i) > 0 \iff s_a(e_i) = 0, \end{cases}$$

so the error rate is

$$\begin{aligned} \mathbb{P}(\mathcal{T}_a \circ y_a(X) \neq Y \mid A = a) &= \sum_{i\in[k]} p_a^*(e_i) \cdot (1 - \mathbb{P}(\mathcal{T}_a \circ y_a(X) = e_i \mid A = a, Y = e_i)) \\ &= \sum_{i\in[k]} p_a^*(e_i)s_a(e_i) \\ &= \sum_{i\in[k]} \max(0, p_a^*(e_i) - q_a^*(e_i)) \\ &= \frac{1}{2} \sum_{i\in[k]} |p_a^*(e_i) - q_a^*(e_i)| = D_{\mathrm{TV}}(p_a^*, q_a^*). \end{aligned}$$

We conclude by averaging the error on all groups and invoking Theorem A.2. □

A.3 Two Examples in Remarks

In the remarks of Theorem 3.2, we discussed properties of the inherent tradeoff of error rate for DP fairness, which we illustrated here with two concrete examples.

It is discussed that the tradeoff could be zero even when the distribution of class probabilities $r_a^* \coloneqq f_a^*\sharp\mu_a^X$ differ, or equivalently, $\mathbb{E}_\mu[Y \mid X, A] \perp\!\!\!\perp A$. This means that on certain problem instances, the Bayes error rate is simultaneously achieved by an unfair classifier and a fair one; in other words, the cost of DP fairness is zero. Such cases arise from the nonuniqueness of the optimal classifier. They would not occur on regression problems (with MSE), where the optimal regressor is always unique (namely, $f_a^*(x) = \mathbb{E}_\mu[Y \mid X, A = a]$).

Example A.5. Consider the two-group binary classification problem given by

$$\mathbb{P}_{\mu_1}(Y = e_1 \mid X = x) = 1 \quad \text{and}$$
$$\mathbb{P}_{\mu_2}(Y = e_1 \mid X = x) = \mathbb{P}_{\mu_2}(Y = e_2 \mid X = x) = \frac{1}{2} \quad \text{for all } x \in \mathcal{X}.$$

The optimal classifier on group 1 is the constant function $x \mapsto e_1$, and all classifiers on group 2 yield the same (hence optimal) error rate of $\frac{1}{2}$, including $x \mapsto e_1$ which when combined with the optimal group 1 classifier achieves DP and the (group-balanced) Bayes error rate of $\frac{1}{4}$.

In (Zhao and Gordon, 2022), it is concluded that in the noiseless setting, the inherent tradeoff is zero if and only if the class marginal distributions are the same, $p^*_a(e_i) = \mathbb{P}_\mu(Y = e_i \mid A = a)$ and $r^*_a = p^*_a \coloneqq f^*_a \sharp \mu^X_a$ in this case, or equivalently $\mathbb{E}_\mu[Y \mid A] \perp\!\!\!\perp A$. However, for the general case, this is no longer sufficient for the tradeoff to be zero (a sufficient condition here is $\mathbb{E}_\mu[Y \mid X, A] \perp\!\!\!\perp A$).

Example A.6. Consider the two-group binary classification problem of two inputs, $\mathcal{X} = \{1, 2\}$, given by

$$\begin{aligned}
&\mathbb{P}_{\mu_1}(Y = e_1 \mid X = 1) = 1, \quad \mathbb{P}_{\mu_1}(Y = e_1 \mid X = 2) = 0, \\
&\mathbb{P}_{\mu_1}(Y = e_2 \mid X = 1) = 0, \quad \mathbb{P}_{\mu_1}(Y = e_2 \mid X = 2) = 1, \quad \text{with} \\
&\qquad \mathbb{P}_{\mu_1}(X = 1) = \frac{1}{3}, \qquad \mathbb{P}_{\mu_1}(X = 2) = \frac{2}{3}, \quad \text{and} \\
&\qquad \mathbb{P}_{\mu_2}(Y = e_1 \mid X = x) = \frac{1}{3} \quad \text{for all } x \in \{1, 2\}.
\end{aligned}$$

Note that the class marginal on both groups is $(\frac{1}{3}, \frac{2}{3})$. The unique optimal classifier on group 1 is $x \mapsto e_x$, and the unique optimal classifier on group 2 is the constant $x \mapsto e_2$, but this combination do not satisfy DP, since the output distribution on group 1 is $(\frac{1}{3}, \frac{2}{3})$ but that on group 2 is $(0, 1)$. Since all other classifiers including the fair ones have strictly higher error rates, the tradeoff is nonzero.

B Omitted Proofs from Section 3

To make our arguments rigorous, we provide a definition of randomized functions via the Markov kernel. These definitions will be frequently referred to in the proofs in this section, and that of Theorem 4.5.

Definition B.1 (Markov Kernel)**.** A Markov kernel from a measurable space $(\mathcal{X}, \mathcal{S})$ to $(\mathcal{Y}, \mathcal{T})$ is a mapping $\mathcal{K} : \mathcal{X} \times \mathcal{T} \to [0, 1]$, such that $\mathcal{K}(\cdot, T)$ is $\mathcal{S}$-measurable $\forall T \in \mathcal{T}$, and $\mathcal{K}(x, \cdot)$ is a probability measure on $(\mathcal{Y}, \mathcal{T})$ $\forall x \in \mathcal{X}$.

Definition B.2 (Randomized Function)**.** A randomized function $f : (\mathcal{X}, \mathcal{S}) \to (\mathcal{Y}, \mathcal{T})$ is associated with a Markov kernel $\mathcal{K} : \mathcal{X} \times \mathcal{T} \to [0, 1]$, and for all $x \in \mathcal{X}, T \in \mathcal{T}$, $\mathbb{P}(f(x) \in T) = \mathcal{K}(x, T)$.

Definition B.3 (Push-Forward by Randomized Function)**.** Let p be a measure on $(\mathcal{X}, \mathcal{S})$ and $f : (\mathcal{X}, \mathcal{S}) \to (\mathcal{Y}, \mathcal{T})$ a randomized function with Markov kernel $\mathcal{K}$. The push-forward of p under f, denoted by $f\sharp p$, is a measure on $\mathcal{Y}$ given by $f\sharp p(T) = \int_{\mathcal{X}} \mathcal{K}(x, T) \, \mathrm{d}p(x)$ for all $T \in \mathcal{T}$.

Also, let blackboard bold $\mathbb{1}$ denote the indicator function, where $\mathbb{1}(E) = 1$ if the predicate E is true, else 0.

We provide the proofs to Lemma 3.1 and Theorems 3.3 and 3.4; a proof of Theorem 3.2 was provided in Appendix A.1. The proofs to these results and the ones in Section 4 all make use of the following rewriting of the error rate as an integral over a coupling:

Lemma B.4. *Let μ with $|\mathcal{A}| = 1$ be given along with the Bayes optimal predictor f^*. Then for any randomized classifier $h : \mathcal{X} \to \mathcal{Y}$, its error rate can be written as*

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

Note that the joint distribution $\mathbb{P}$ of $(f^*(X), h(X))$ is a coupling belonging to $\Gamma(f^*\sharp\mu^X, h\sharp\mu^X)$.

Proof. The accuracy of h is

$$\begin{aligned}
1 - \operatorname{err}(h) &= 1 - \mathbb{P}(h(X) \neq Y) = \mathbb{P}(h(X) = Y) \\
&= \int_{\Delta_k} \sum_{i\in[k]} \mathbb{P}(Y = e_i, h(X) = e_i, f^*(X) = s)\,\mathrm{d}s \\
&= \int_{\Delta_k} \sum_{i\in[k]} \mathbb{P}(Y = e_i, h(X) = e_i \mid f^*(X) = s) \cdot \mathbb{P}_\mu(f^*(X) = s)\,\mathrm{d}s \\
&= \int_{\Delta_k} \sum_{i\in[k]} \mathbb{P}_\mu(Y = e_i \mid f^*(X) = s) \cdot \mathbb{P}(h(X) = e_i \mid f^*(X) = s) \cdot \mathbb{P}_\mu(f^*(X) = s)\,\mathrm{d}s \\
&= \int_{\Delta_k} \sum_{i\in[k]} s_i \cdot \mathbb{P}(f^*(X) = s, h(X) = e_i)\,\mathrm{d}s,
\end{aligned}$$

where line 4 follows from $X \perp\!\!\!\perp Y$ given $f^*(X)$, since $f^*(X) = \mathbb{E}_\mu[Y \mid X]$ fully specifies the pmf of Y conditioned on X. Next, because $\mathbb{P}(f^*(X) = \cdot, h(X) = \cdot)$ is a probability measure,

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

where the second equality is due to an identity stated in Eq. (10) between points in the simplex and its vertices. □

Lemma B.5 (Full Version of Lemma 3.1)**.** *Let μ with $|\mathcal{A}| = 1$ be given along with the Bayes optimal predictor f^*, define $r^* \coloneqq f^*\sharp\mu^X$, and fix $q \in \mathcal{Q}_k$. Then for any randomized classifier $h : \mathcal{X} \to \mathcal{Y}$ with Markov kernel $\mathcal{K}$ satisfying $h\sharp\mu^X = q$, the coupling $\gamma \in \Gamma(r^*, q)$ given by*

$$\gamma(s, y) = \int_{{f^*}^{-1}(s)} \mathcal{K}(x, y)\,\mathrm{d}\mu^X(x)$$

satisfies

$$\operatorname{err}(h) = \frac{1}{2}\int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1\,\mathrm{d}\gamma(s, y). \tag{8}$$

Conversely, for any $\gamma \in \Gamma(r^*, q)$, the randomized classifier h with Markov kernel

$$\mathcal{K}(x, T) = \gamma(f^*(x), T)/\gamma(f^*(x), \mathcal{Y})$$

satisfies $h\sharp\mu^X = q$ and Eq. (8).

Proof. We begin with the first direction. Let a randomized classifier h with Markov kernel $\mathcal{K}$ satisfying $h\sharp\mu^X = q$ be given. We verify that the coupling constructed above belongs to $\Gamma(r^*, q)$:

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

where line 3 follows from Definition B.1 of Markov kernels, and line 5 from the definition of push-forward measures;

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

where line 3 follows from Definition B.2 of randomized function, and line 5 is by assumption.

Next, by Lemma B.4 and the same arguments above,

$$\begin{aligned}
\mathrm{err}(h) &= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \cdot \mathbb{P}(f^*(X) = s, h(X) = y)\,\mathrm{d}(s,y) \\
&= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \left(\int_x \mathbb{P}(f^*(X) = s, h(X) = y, X = x)\,\mathrm{d}x \right) \mathrm{d}(s,y) \\
&= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \left(\int_{f^{*-1}(s)} \mathbb{P}(h(X) = y, X = x)\,\mathrm{d}x \right) \mathrm{d}(s,y) \\
&= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \left(\int_{f^{*-1}(s)} \mathbb{P}(h(X) = y \mid X = x)\,\mathrm{d}\mu^X(x) \right) \mathrm{d}(s,y) \\
&= \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \cdot \gamma(s,y)\,\mathrm{d}(s,y)
\end{aligned}$$

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

For the converse, let a coupling $\gamma \in \Gamma(r^*, q)$ be given. We show that the Markov kernel of the randomized classifier h constructed in the statement satisfies the equality $\gamma(s,y) = \int_{f^{*-1}(s)} \mathcal{K}(x,y)\,\mathrm{d}\mu^X(x)$, then Eq. (8) will follow directly from the same arguments used in the previous part. Let $s \in \Delta_k$

and $y \in \mathcal{Y}$, and $x' \in f^{*-1}(s)$ arbitrary, then

$$\begin{aligned}
\gamma(s,y) &= \frac{\gamma(s,y)}{\gamma(s,\mathcal{Y})} \cdot \gamma(s,\mathcal{Y}) \\
&= \frac{\gamma(f^*(x'),y)}{\gamma(f^*(x'),\mathcal{Y})} \cdot \gamma(s,\mathcal{Y}) \\
&= \mathcal{K}(x',y) \cdot \gamma(s,\mathcal{Y}) \\
&= \mathcal{K}(x',y) \cdot r^*(s) \\
&= \mathcal{K}(x',y) \int_{x \in f^{*-1}(s)} \mathrm{d}\mu^X(x) \\
&= \int_{x \in f^{*-1}(s)} \mathcal{K}(x',y)\, \mathrm{d}\mu^X(x) \\
&= \int_{x \in f^{*-1}(s)} \mathcal{K}(x,y)\, \mathrm{d}\mu^X(x),
\end{aligned}$$

where line 3 is by construction of $\mathcal{K}$, line 4 from $\gamma \in \Gamma(r^*, q)$, and the last line is because $\mathcal{K}(x,y)$ is constant for all $x \in f^{*-1}(s)$, also by construction. □

Proof of Theorem 3.3. By construction, the Markov kernel of the randomized optimal fair classifier $\bar{h}^*(x,a) \coloneqq \mathcal{T}^*_{r^*_a \to q^*_a} \circ f^*_a(x)$ is

$$\mathcal{K}((x,a),y) = \frac{\gamma^*_a(f^*_a(x),y)}{\gamma^*_a(f^*_a(x),\mathcal{Y})}$$

where $\gamma^*_a \in \Gamma(r^*_a, q^*_a)$ is the optimal transport between r^*_a and q^*_a.

We verify that the output distributions of $\bar{h}^*$ equal $q^*_1, \cdots, q^*_m$, thereby it is ϵ-DP because the q^*_a's satisfy the constraint in Eq. (6), and its error rate achieves the minimum in Theorem 3.2.

First, for all $y \in \mathcal{Y}$,

$$\begin{aligned}
\mathbb{P}(\bar{h}^*(X,A) = y \mid A = a) &= \int_{\mathcal{X}} \mathbb{P}(\bar{h}^*(x,a) = y) \cdot \mathbb{P}_\mu(X = x \mid A = a)\, \mathrm{d}x \\
&= \int_{\mathcal{X}} \mathcal{K}((x,a),y)\, \mathrm{d}\mu^X_a(x) \\
&= \int_{\mathcal{X}} \frac{\gamma^*_a(f^*_a(x),y)}{\gamma^*_a(f^*_a(x),\mathcal{Y})}\, \mathrm{d}\mu^X_a(x) \\
&= \int_{\Delta_k} \frac{\gamma^*_a(s,y)}{\gamma^*_a(s,\mathcal{Y})} \left(\int_{f^{*-1}_a(s)} \mathrm{d}\mu^X_a(x) \right) \mathrm{d}s \\
&= \int_{\Delta_k} \frac{\gamma^*_a(s,y)}{\gamma^*_a(s,\mathcal{Y})} r^*_a(s)\, \mathrm{d}s \\
&= \int_{\Delta_k} \gamma^*_a(s,y)\, \mathrm{d}s = \gamma^*_a(\Delta_k, y) = q^*_a(y),
\end{aligned}$$

where line 3 is due to $\gamma^*_a(f^*_a(x),y) = \gamma^*_a(s,y)$ being constant for all $x \in f^{*-1}_a(s)$.

Similarly, Lemma B.5 implies that the error rate on group a, denoted by $\mathrm{err}_a(\bar{h}^*)$, is

$$\mathrm{err}_a(\bar{h}^*) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1\, \mathrm{d}\gamma_a(s,y),$$

where $\gamma_a \in \Gamma(r_a^*, q_a^*)$ equals to

$$\begin{aligned}\gamma_a(s,y) &= \int_{{f_a^*}^{-1}(s)} \mathcal{K}((x,a),y)\,\mathrm{d}\mu_a^X(x)\\ &= \int_{{f_a^*}^{-1}(s)} \frac{\gamma_a^*(f_a^*(x),y)}{\gamma_a^*(f_a^*(x),\mathcal{Y})}\,\mathrm{d}\mu_a^X(x)\\ &= \frac{\gamma_a^*(s,y)}{\gamma_a^*(s,\mathcal{Y})}\int_{{f_a^*}^{-1}(s)} \mathrm{d}\mu_a^X(x)\\ &= \gamma_a^*(s,y).\end{aligned}$$

So $\mathrm{err}_a(\bar{h}^*) = \frac{1}{2}\int_{\Delta_k\times\mathcal{Y}} \|s-y\|_1\,\mathrm{d}\gamma_a^*(s,y) = \frac{1}{2}W_1(r_a^*, q_a^*)$ because γ_a^* is an optimal transport between r_a^* and q_a^*, and $\mathrm{err}(\bar{h}^*) = \frac{1}{m}\sum_{a\in[m]}\mathrm{err}_a(\bar{h}^*) = \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(r_a^*, q_a^*)$, the minimum error rate under ϵ-DP. □

Proof of Theorem 3.4. The ϵ-fair classifier $\bar{h}$ returned from Algorithm 1 on $f_1, \cdots, f_m$ is

$$\bar{h}(x,a) = \mathcal{T}^*_{r_a\to q_a} \circ f_a(x),$$

where $q_1, \cdots, q_m$ is the minimizer of Eq. (6) on $r_a \coloneqq f_a\sharp\mu_a^X$ and ϵ, and $\mathcal{T}^*_{r_a\to q_a}$ is the optimal transport from r_a to q_a.

Denote the L^1 error of f_a relative to f_a^* conditioned on $A = a$ by

$$\mathcal{E}_a \coloneqq \mathbb{E}_{\mu_a^X}[\|f_a(X) - f_a^*(X)\|_1],$$

and write $\mathcal{E} \coloneqq \frac{1}{m}\sum_{a\in[m]}\mathcal{E}_a$.

For the upper bound, by Lemma B.4 and the triangle inequality,

$$\begin{aligned}\mathrm{err}(\bar{h}) &= \frac{1}{2m}\sum_{a\in[m]}\mathbb{E}\big[\|\mathcal{T}^*_{r_a\to q_a}\circ f_a(X) - f_a^*(X)\|_1 \mid A = a\big]\\ &\le \frac{1}{2m}\sum_{a\in[m]}\big(\mathbb{E}[\|\mathcal{T}^*_{r_a\to q_a}\circ f_a(X) - f_a(X)\|_1 \mid A=a] + \mathbb{E}[\|f_a(X) - f_a^*(X)\|_1 \mid A=a]\big)\\ &= \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(r_a, q_a) + \frac{1}{2}\cdot\frac{1}{m}\sum_{a\in[m]}\mathcal{E}_a\\ &= \min_{\substack{q_1',\cdots,q_m'\in\mathcal{Q}_k\\ \max_{a,a'} D_{\mathrm{TV}}(q_a', q_{a'}')\le\epsilon}} \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(r_a, q_a') + \frac{\mathcal{E}}{2},\end{aligned}$$

where line 3 is because $\mathcal{T}^*_{r_a\to q_a}$ is the optimal transport from r_a to q_a under the ℓ_1 cost, and line 4 because $q_1, \cdots, q_m$ is the minimizer. Let $q_1^*, \cdots, q_m^*$ denote the minimizer of Eq. (6) on the $r_a^* \coloneqq f_a^*\sharp\mu_a^X$ and ϵ, then by Theorem 3.2,

$$\begin{aligned}\mathrm{err}(\bar{h}) - \mathrm{err}_\epsilon^* &\le \frac{1}{2m}\left(\min_{\substack{q_1',\cdots,q_m'\in\mathcal{Q}_k\\ \max_{a,a'} D_{\mathrm{TV}}(q_a', q_{a'}')\le\epsilon}} \sum_{a\in[m]} W_1(r_a, q_a') - \sum_{a\in[m]} W_1(r_a^*, q_a^*)\right) + \frac{\mathcal{E}}{2}\\ &\le \frac{1}{2m}\left(\sum_{a\in[m]} W_1(r_a, q_a^*) - \sum_{a\in[m]} W_1(r_a^*, q_a^*)\right) + \frac{\mathcal{E}}{2}\\ &\le \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(r_a, r_a^*) + \frac{\mathcal{E}}{2} \le \frac{1}{2}\cdot\frac{1}{m}\sum_{a\in[m]}\mathcal{E}_a + \frac{\mathcal{E}}{2} = \mathcal{E},\end{aligned}$$

where the last line is because for each $a \in [m]$, $W_1(r_a, r_a^*)$ is upper bounded by the transportation cost under the coupling given by the joint distribution of $(f_a(X), f_a^*(X))$ conditioned on $A = a$: denote the coupling by π_a, then clearly $\pi_a \in \Gamma(r_a, r_a^*)$, and $\int_{\Delta_k \times \Delta_k} \|s - s'\|_1 \,\mathrm{d}\pi_a(s, s') = \int_{\mathcal{X}} \|f_a(x) - f_a^*(x)\|_1 \,\mathrm{d}\mu_a^X(x) = \mathcal{E}_a$.

For the lower bound, again by Lemma B.4,

$$\begin{aligned}\mathrm{err}(\bar{h}) &= \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \cdot \mathbb{P}(f_a^*(X) = s, \mathcal{T}_{r_a \to q_a}^* \circ f_a(X) = y \mid A = a) \,\mathrm{d}(s, y) \\ &\geq \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} W_1(r_a^*, q_a) \geq \min_{\substack{q_1', \cdots, q_m' \in \mathcal{Q}_k \\ \max_{a,a'} D_{\mathrm{TV}}(q_a', q_{a'}') \leq \epsilon}} \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} W_1(r_a^*, q_a') = \mathrm{err}_\epsilon^*,\end{aligned}$$

where line 2 is because the joint distribution of $(f_a^*(X), \mathcal{T}_{r_a \to q_a}^* \circ f_a(X))$ conditioned on $A = a$ is a coupling belonging to $\Gamma(r_a^*, q_a)$, thereby the transportation cost represented by the quantity in the preceding line upper bounds $W_1(r_a^*, q_a)$. □

C Omitted Proofs from Section 4

We provide proofs to the generalization results (Theorems 4.3 and 4.4) and the error propagation bound for smoothing (Theorem 4.5), in that order. We remark that Assumption 4.2 of the predictors to be post-processed being calibrated can be dropped by adding the error propagation of Theorem 3.4 (the L^1 error w.r.t. the calibrated versions; via combining the proofs). A finite sample generalization bound for the general case procedure in Section 4.3 can also be obtained by combining Theorems 4.4 and 4.5, provided that the supports of the distributions after smoothing are contained in the simplex.

The generalization bound of the finite case uses an ℓ_1 (TV) convergence result of empirical distributions, which is obtained from the following vector concentration bound (Kohler and Lucchi, 2017, Lemma 18):

Lemma C.1 (Vector Bernstein Inequality)**.** *Let $x_1, \cdots, x_n \in \mathbb{R}^d$ be independent vector-valued random variables, satisfying $\mathbb{E}[x_i] = 0$, $\|x_i\|_2 \leq \mu$, and $\mathbb{E}[\|x_i\|_2^2] \leq \sigma^2$ for all $i \in [n]$. Then for all $\epsilon \in (0, \sigma^2/\mu)$,*

$$\mathbb{P}\left(\left\| \frac{1}{n} \sum_{i=1}^n x_i \right\|_2 \geq \epsilon \right) \leq \exp\left(-\frac{n\epsilon^2}{8\sigma^2} + \frac{1}{4} \right).$$

Lemma C.2. *Let p be a distribution over $\mathcal{X}$ with finite support, and $x_1, \cdots, x_n \sim p$ be i.i.d. samples. Define the empirical distribution $\hat{p}_n \coloneqq \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Then w.p. at least $1 - \delta$ over the random draw of the samples, $\|p - \hat{p}_n\|_1 \leq \sqrt{\frac{32|\mathcal{X}|}{n}\left(\frac{1}{4} + \ln\left(\frac{1}{\delta}\right)\right)}$.*

Proof. By viewing each δ_{x_i} as a vector in $\Delta_{|\mathcal{X}|}$, we have that $p - \delta_{x_i}$ is contained in the $|\mathcal{X}|$-dimensional ℓ_1-ball of width 2, whereby $\|p - \delta_{x_i}\|_2^2 \leq \|p - \delta_{x_i}\|_1^2 \leq 4$ for all i. So by Lemma C.1, w.p. at least $1 - \delta$, $\|p - \frac{1}{n} \sum_{i=1}^n \delta_{x_i}\|_1 \leq \sqrt{|\mathcal{X}|} \|p - \frac{1}{n} \sum_{i=1}^n \delta_{x_i}\|_2 \leq \sqrt{\frac{32|\mathcal{X}|}{n}\left(\frac{1}{4} + \ln\left(\frac{1}{\delta}\right)\right)}$. □

Proof of Theorem 4.3. We first bound the error rate, then the fairness.

Error Rate. Consider the classification problem μ' derived from the original μ under an input transformation given by the joint distribution of $(f_A(X), A, Y)$, as discussed in Section 3.3, on which

Id is the Bayes optimal predictor due to calibration of the f_a's. Then by Lemma B.4 applied on μ',

$$\begin{aligned}
\mathrm{err}(\hat{h}) &= \frac{1}{2m}\sum_{a\in[m]}\sum_{s\in\mathcal{R}_a}\sum_{y\in\mathcal{Y}}\|s-y\|_1\cdot\mathbb{P}(\mathrm{Id}(X')=s,\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(X')=y)\\
&= \frac{1}{2m}\sum_{a\in[m]}\sum_{s\in\mathcal{R}_a}\sum_{y\in\mathcal{Y}}\|s-y\|_1\cdot r_a(s)\cdot\mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s)=y)\\
&\le \frac{1}{2m}\sum_{a\in[m]}\sum_{s\in\mathcal{R}_a}\sum_{y\in\mathcal{Y}}\|s-y\|_1\cdot(\hat{r}_a(s)+|r_a(s)-\hat{r}_a(s)|)\cdot\mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s)=y)\\
&\le \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(\hat{r}_a,\hat{q}_a)+\frac{1}{m}\sum_{a\in[m]}\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|,
\end{aligned}$$

where line 4 uses the fact that each $\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}$ is an optimal transport from $\hat{r}_a$ to $\hat{q}_a$. Let the q_a's denote the minimizer of Eq. (6) on the r_a's with ϵ, then

$$\begin{aligned}
\mathrm{err}(\hat{h})-\mathrm{err}^*_{\epsilon,f} &\le \frac{1}{m}\sum_{a\in[m]}\left(\frac{1}{2}(W_1(\hat{r}_a,\hat{q}_a)-W_1(r_a,q_a))+\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\right)\\
&\le \frac{1}{m}\sum_{a\in[m]}\left(\frac{1}{2}(W_1(\hat{r}_a,q_a)-W_1(r_a,q_a))+\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\right)\\
&\le \frac{1}{m}\sum_{a\in[m]}\left(\frac{1}{2}W_1(\hat{r}_a,r_a)+\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\right),
\end{aligned}$$

where line 2 is due to $\hat{q}_a$ being the minimizer of Eq. (6) on the $\hat{r}_a$'s. Because $\|s-s'\|_1\le 2\cdot\mathbb{1}(s\neq s')$, by Proposition A.3, $W_1(\hat{r}_a,r_a)\le 2D_{\mathrm{TV}}(\hat{r}_a,r_a)$, so it follows that

$$\mathrm{err}(\hat{h})-\mathrm{err}^*_{\epsilon,f}\le\frac{3}{2m}\sum_{a\in[m]}\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\le\frac{1}{m}\sum_{a\in[m]}\sqrt{\frac{72|\mathcal{R}_a|}{n}\left(\frac{1}{4}+\ln\Big(\frac{m}{\delta}\Big)\right)}$$

w.p. at least $1-\delta$ from m applications of Lemma C.2 and a union bound.

Fairness (Δ_{DP}). Note that for all $a\in[m]$ and $y\in\mathcal{Y}$,

$$\begin{aligned}
&\sum_{y\in\mathcal{Y}}\Big|\mathbb{P}(\hat{h}(X,A)=y\mid A=a)-\hat{q}_a(y)\Big|\\
&\quad=\sum_{y\in\mathcal{Y}}\left|\sum_{s\in\mathcal{R}_a}r_a(s)\cdot\mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s)=y)-\sum_{s\in\mathcal{R}_a}\hat{r}_a(s)\cdot\mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s)=y)\right|\\
&\quad\le\sum_{y\in\mathcal{Y}}\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\cdot\mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s)=y)\\
&\quad=\sum_{s\in\mathcal{R}_a}|r_a(s)-\hat{r}_a(s)|\\
&\quad\le\sqrt{\frac{32|\mathcal{R}_a|}{n}\left(\frac{1}{4}+\ln\Big(\frac{m}{\delta}\Big)\right)}
\end{aligned}$$

w.p. at least $1-\delta$ from applications of Lemma C.2 and a union bound, where the first equality is because $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ is a transport from $\hat{r}_a$ to $\hat{q}_a$. Then, because of the constraint in Eq. (6) that $\max_{a,a' \in [m]} D_{\mathrm{TV}}(\hat{q}_a, \hat{q}_{a'}) \le \epsilon$,

$$
\begin{aligned}
&\Delta_{\mathrm{DP}}(\hat{h}) \\
&\quad = \max_{a,a' \in [m]} \frac{1}{2} \sum_{y \in \mathcal{Y}} \Big| \mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \mathbb{P}(\hat{h}(X,a') = y \mid A = a') \Big| \\
&\quad \le \max_{a,a' \in [m]} \frac{1}{2} \sum_{y \in \mathcal{Y}} |\hat{q}_a(y) - \hat{q}_{a'}(y)| \\
&\qquad + \max_{a,a' \in [m]} \frac{1}{2} \sum_{y \in \mathcal{Y}} \Big(\Big| \mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y) \Big| + \Big| \mathbb{P}(\hat{h}(X,a') = y \mid A = a') - \hat{q}_{a'}(y) \Big| \Big) \\
&\quad \le \epsilon + \max_{a \in [m]} \sqrt{\frac{32 |\mathcal{R}_a|}{n} \Big(\frac{1}{4} + \ln\Big(\frac{m}{\delta} \Big) \Big)}.
\end{aligned}
$$

The theorem then follows from a final application of union bound. □

The proof of the generalization bound in the continuous case uses the following uniform bounds with the VC dimension and the pseudo-dimension as the complexity measure. We omit the proofs, but refer readers to (Shalev-Shwartz and Ben-David, 2014, Theorem 6.8) and (Mohri et al., 2018, Theorem 11.8), respectively. We also need a characterization of the disagreement between the empirical transports $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$ and the transport mappings $\mathcal{T}_a \in \mathcal{G}_k$ extracted from them in Lines 6–9 of Algorithm 2, (to be) stated in Lemma D.5.

Theorem C.3. *Let a class of binary functions $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ be given, and $\ell(\hat{y}, y) = \mathbb{1}(y \neq y)$ be the 0-1 loss. Let p be a distribution over $\mathcal{X} \times \{0,1\}$, and $(x_1, y_1), \cdots, (x_n, y_n) \sim p$ be i.i.d. samples. Then w.p. at least $1-\delta$ over the random draw of the samples, $\forall h \in \mathcal{H}$,*

$$
\left| \mathbb{E}_{(X,Y) \sim p}[\ell(h(X), Y)] - \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \right| \le \sqrt{\frac{C}{n} \Big(d + \ln\Big(\frac{1}{\delta} \Big) \Big)},
$$

where $C > 0$ is a universal constant, and d is the VC-dimension of $\mathcal{H}$.

Theorem C.4. *Let a class of real-valued functions $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be given, along with a non-negative loss function ℓ that is upper bounded by B. Let p be a distribution over $\mathcal{X} \times \mathbb{R}$, and $(x_1, y_1), \cdots, (x_n, y_n) \sim p$ be i.i.d. samples. Then w.p. at least $1-\delta$ over the random draw of the samples, $\forall h \in \mathcal{H}$,*

$$
\mathbb{E}_{(X,Y) \sim p}[\ell(h(X), Y)] - \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) \le O\left(B \sqrt{\frac{1}{n} \Big(d \ln\Big(\frac{n}{d} \Big) + \ln\Big(\frac{1}{\delta} \Big) \Big)} \right),
$$

where d is the pseudo-dimension of $\{(x,y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$, the class of ℓ associated to $\mathcal{H}$.

One highlight of our proof is that we avoided using the convergence of the empirical measure under Wasserstein distance in our arguments, which would have resulted in sample complexity that is exponential in the number of label classes k (Weed and Bach, 2019). Instead, we leveraged the existence and uniqueness of the semi-discrete simplex-vertex optimal transport in the low complexity function class $\mathcal{G}_k$, established in Theorems D.1 and D.2, whereby we can apply the above uniform bound to $\mathcal{G}_k$ and achieve a rate that is only polynomial in k.

In addition, we remark that the $O(k^2/n)$ term attributed to the disagreements between $\mathcal{T}_a$ and $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ on S_a could potentially be improved to $O(k/n)$ if OPT is assumed to return an extremal solution (Peyré and Cuturi, 2019), and $\mathcal{G}_k$ is modified so that the output on points that lie on each boundary can be specified, rather than always tie-broken to the e_i with the largest index i. Also, the $\widetilde{O}(\sqrt{k^3/n})$ sample complexity for Δ_{DP} could be improved using more elaborate generalization analyses, instead of off-the-shelf VC bounds that we applied coordinate-wise.

Proof of Theorem 4.4. We first bound the error rate, followed by Δ_{DP}. Recall that the classifier returned from Algorithm 2 is

$$\hat{h}(x, a) \coloneqq \mathcal{T}_a \circ f_a(x),$$

where each $\mathcal{T}_a \in \mathcal{G}_k$ is extracted from the empirical optimal transport $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ by Algorithm 2, obtained from calling $\mathrm{OPT}(\hat{r}_1, \cdots, \hat{r}_m, \epsilon)$, where $\hat{q}_1, \cdots, \hat{q}_m$ is the minimizer of Eq. (6) on the $\hat{r}_a$'s with ϵ. We will use a complexity result of $\mathcal{G}_k$ in terms of its pseudo-dimension when associated with ℓ_1 loss, and a VC bound of binarized versions of $\mathcal{G}_k$ (to be defined in Eq. (19)), which are deferred to Theorem D.15 and Corollary D.16.

Error Rate. Consider the classification problem μ' derived from the original μ under an input transformation given by the joint distribution of $(f_A(X), A, Y)$, as discussed in Section 3.3, on which Id is the Bayes optimal predictor due to calibration of the f_a's. Then by Lemma B.4 applied on μ',

$$\begin{aligned}
\mathrm{err}(\hat{h}) &= \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \cdot \mathbb{P}(\mathrm{Id}(X') = s, \mathcal{T}_a(X') = y) \, \mathrm{d}(s, y) \\
&= \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} \mathbb{E}_{S \sim r_a}[\|\mathcal{T}_a(S) - S\|_1].
\end{aligned}$$

Define $s_{a,j} \coloneqq f_a(x_{a,j})$. By Theorems C.4 and D.15, and a union bound, we have w.p. at least $1 - \delta$, for all $a \in [m]$,

$$\mathbb{E}_{S \sim r_a}[\|\mathcal{T}_a(S) - S\|_1] - \frac{1}{n} \sum_{j=1}^{n} \|\mathcal{T}_a(s_{a,j}) - s_{a,j}\|_1 \le O\left(\sqrt{\frac{1}{n} \left(k \ln\left(\frac{n}{k}\right) + \ln\left(\frac{m}{\delta}\right) \right)} \right).$$

Because each $\mathcal{T}_a$ is extracted from the empirical optimal transport $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ using Line 6–9 of Algorithm 2, by the discussion in Appendix D and Lemma D.5, they both agree on all of S_a except for points that lie on the decision boundaries of $\mathcal{T}_a$. The boundaries are described by $\binom{k}{2}$ hyperplanes, and because r_a is continuous, no two points in S_a lie on the same hyperplane almost surely, so the number of disagreements is at most $\binom{k}{2} = k(k-1)/2$, and

$$\begin{aligned}
\left| \frac{1}{n} \sum_{j=1}^{n} \|\mathcal{T}_a(s_{a,j}) - s_{a,j}\|_1 - W_1(\hat{r}_a, \hat{q}_a) \right| &= \left| \frac{1}{n} \sum_{j=1}^{n} \|\mathcal{T}_a(s_{a,j}) - s_{a,j}\|_1 - \frac{1}{n} \sum_{j=1}^{n} \|\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}(s_{a,j}) - s_{a,j}\|_1 \right| \\
&\le \frac{1}{n} \sum_{j=1}^{n} \|\mathcal{T}_a(s_{a,j}) - \mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}(s_{a,j})\|_1 \\
&= \frac{2}{n} \sum_{j=1}^{n} \mathbb{1}(\mathcal{T}_a(s_{a,j}) \neq \mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}(s_{a,j})) \le O\left(\frac{k^2}{n}\right),
\end{aligned}$$

where the first equality is because $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ is the optimal transport from $\hat{r}_a$ to $\hat{q}_a$. Therefore, we arrive at w.p. at least $1-\delta$,

$$\mathrm{err}(\hat{h}) - \frac{1}{m}\sum_{a\in[m]} \frac{1}{2} W_1(\hat{r}_a, \hat{q}_a) \le O\left(\sqrt{\frac{1}{n}\Big(k\ln\Big(\frac{n}{k}\Big) + \ln\Big(\frac{m}{\delta}\Big)\Big)} + \frac{k^2}{n}\right) =: \mathcal{E}.$$

Continuing, let $\mathcal{T}^*_{r_a \to q_a} \in \mathcal{G}_k$ denote the optimal transport from r_a to q_a, where the q_a's denote the minimizer of Eq. (6) on the r_a's with ϵ. The existence of this transport in $\mathcal{G}_k$ is due to the problem being semi-discrete and Theorem D.1. Define $q'_a(y) = \frac{1}{n}\sum_{j=1}^n \mathbb{1}(\mathcal{T}^*_{r_a\to q_a}(s_{a,j}) = y)$. It follows that

$$\begin{aligned}
\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\epsilon,f} &\le \mathcal{E} + \frac{1}{m}\sum_{a\in[m]} \frac{1}{2}(W_1(\hat{r}_a, \hat{q}_a) - W_1(r_a, q_a)) \\
&= \mathcal{E} + \frac{1}{m}\sum_{a\in[m]} \frac{1}{2}\big(W_1(\hat{r}_a, \hat{q}_a) - W_1(\hat{r}_a, q_a) + W_1(\hat{r}_a, q_a) - W_1(\hat{r}_a, q'_a) \\
&\qquad\qquad + W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a)\big) \\
&\le \mathcal{E} + \frac{1}{m}\sum_{a\in[m]} \frac{1}{2}\Big((W_1(\hat{r}_a, q_a) - W_1(\hat{r}_a, q'_a)) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a))\Big) \\
&\le \mathcal{E} + \frac{1}{m}\sum_{a\in[m]} \frac{1}{2}\Big(W_1(q_a, q'_a) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a))\Big),
\end{aligned}$$

where line 3 is due to $\hat{q}_a$ being the minimizer of Eq. (6) on the $\hat{r}_a$'s

For the first term in the summand, because both distributions q_a, q'_a are supported on the vertices, so by Proposition A.3, $\frac{1}{2}W_1(q_a, q'_a) = D_{\mathrm{TV}}(q_a, q'_a)$, and w.p. at least $1-\delta$, for all $a \in [m]$,

$$\begin{aligned}
D_{\mathrm{TV}}(q_a, q'_a) &= \sum_{i\in[k]} \left| \mathbb{E}_{S\sim r_a}[\mathbb{1}(\mathcal{T}^*_{r_a\to q_a}(S)_i = 1)] - \frac{1}{n}\sum_{j=1}^n \mathbb{1}(\mathcal{T}^*_{r_a\to q_a}(s_{a,j})_i = 1)\right| \\
&= \left\| \mathbb{E}_{S\sim r_a}[\mathcal{T}^*_{r_a\to q_a}(S)] - \frac{1}{n}\sum_{j=1}^n \mathcal{T}^*_{r_a\to q_a}(s_{a,j})\right\|_1 \\
&\le O\left(\sqrt{\frac{k}{n}\ln\Big(\frac{m}{\delta}\Big)}\right),
\end{aligned}$$

where the last line follows from applying Lemma C.1 using the same arguments in the proof of Lemma C.2, and a union bound over all $a \in [m]$.

For the second term, because then the joint probability of $(\widehat{S}, \mathcal{T}^*_{r_a\to q_a}(\widehat{S}))$, $\widehat{S} \sim \hat{r}_a$, is a coupling belonging to $\Gamma(\hat{r}_a, q'_a)$, the transportation cost of $\mathcal{T}^*_{r_a\to q_a}$ on $\hat{r}_a$ to q'_a upper bounds $W_1(\hat{r}_a, q'_a)$, whereby w.p. at least $1-\delta$, for all $a \in [m]$,

$$\begin{aligned}
W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a) &\le \frac{1}{n}\sum_{j=1}^n \|\mathcal{T}^*_{r_a\to q_a}(s_{a,j}) - s_{a,j}\|_1 - \mathbb{E}_{S\sim r_a}[\|\mathcal{T}^*_{r_a\to q_a}(S) - S\|_1] \\
&\le O\left(\sqrt{\frac{1}{n}\ln\Big(\frac{m}{\delta}\Big)}\right)
\end{aligned}$$

by Hoeffding's inequality, because $\|\mathcal{T}^*_{r_a\to q_a}(s_{a,j}) - s_{a,j}\|_1 \le 2$.

Hence, putting everything together, we conclude with a union bound that

$$\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\epsilon,f} \le O\left(\sqrt{\frac{k}{n}\ln\left(\frac{mn}{k\delta}\right)} + \frac{k^2}{n}\right)$$

(assuming $n \ge k$, otherwise $O(\sqrt{k/n \cdot \ln(mn/\delta)} + k^2/n)$).

Fairness (Δ_{DP}). By applying Theorem C.3 and Corollary D.16 to the artificial binary classification problem whose data distribution is the joint distribution of $(S, 1)$, $S \sim r_a$, and a union bound, w.p. at least $1-\delta$, for all $i \in [k]$ and $a \in [m]$,

$$\begin{aligned}&\left|\mathbb{P}(\hat{h}(X,A) = e_i \mid A = a) - \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}(\mathcal{T}_a(s_{a,j}) = e_i)\right| \\ &\quad = \left|\mathbb{E}_{S\sim r_a, \mathcal{T}_a}[\mathbb{1}(\mathcal{T}_a(S)_i = 1)] - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(\mathcal{T}_a(s_{a,j})_i = 1)\right| \le O\left(\sqrt{\frac{k}{n}\ln\left(\frac{mk}{\delta}\right)}\right).\end{aligned}$$

Now, the decision boundaries of the function $s \mapsto \mathbb{1}(\mathcal{T}_a(s)_i = 1) \in \mathcal{G}_{k,i}$ (defined in Eq. (19)) are described by k hyperplanes, and $\mathcal{T}_a$ is extracted from $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$, so by the discussion in Appendix D and Lemma D.5 and the same reasoning used previously, they both agree on all but k points in S_a almost surely, thereby

$$\begin{aligned}\left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(\mathcal{T}_a(s_{a,j}) = e_i) - \hat{q}_a(e_i)\right| &= \left|\frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(\mathcal{T}_a(s_{a,j})_i = 1) - \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s_{a,j})_i = 1)\right| \\ &\le \frac{1}{n}\sum_{j=1}^{n}\mathbb{1}(\mathcal{T}_a(s_{a,j})_i \ne \mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s_{a,j})_i) \le O\left(\frac{k}{n}\right).\end{aligned}$$

Therefore, we conclude that

$$\begin{aligned}&\Delta_{\mathrm{DP}}(\hat{h}) \\ &\quad = \max_{a,a'\in[m]}\frac{1}{2}\sum_{y\in\mathcal{Y}}\left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \mathbb{P}(\hat{h}(X,a') = y \mid A = a')\right| \\ &\quad \le \max_{a,a'\in[m]}\frac{1}{2}\sum_{y\in\mathcal{Y}}|\hat{q}_a(y) - \hat{q}_{a'}(y)| \\ &\qquad + \max_{a,a'\in[m]}\frac{1}{2}\sum_{y\in\mathcal{Y}}\left(\left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y)\right| + \left|\mathbb{P}(\hat{h}(X,a') = y \mid A = a') - \hat{q}_{a'}(y)\right|\right) \\ &\quad \le \epsilon + O\left(\sqrt{\frac{k^3}{n}\ln\left(\frac{mk}{\delta}\right)} + \frac{k^2}{n}\right).\end{aligned}$$

□

Proof of Theorem 4.5. Let the q_a's denote the minimizer of Eq. (6) on the r_a's with ϵ, then

$$\bar{h}_\rho(x,a) \coloneqq \mathcal{T}^*_{\tilde{r}_a\to q_a} \circ u_\rho \circ f_a(x)$$

where $\tilde{r}_a \coloneqq u_\rho \sharp r_a$.

Denote the coupling associated with $\mathcal{T}^*_{\tilde{r}_a \to q_a}$ by $\gamma_a \in \Gamma(\tilde{r}_a, q_a)$, then the Markov kernel of $\mathcal{T}^*_{\tilde{r}_a \to q_a} \circ u_\rho$ is

$$\mathcal{K}(s,T) = \mathbb{E}_{N\sim\rho_s}\left[\frac{\gamma_a(s+N,T)}{\gamma_a(s+N,\mathcal{Y})}\right] = \int_{\tilde{s}\in\mathbb{R}^k} \frac{\gamma_a(\tilde{s},T)}{\gamma_a(\tilde{s},\mathcal{Y})}\,\mathrm{d}(\rho_s * \delta_s)(\tilde{s}) = \int_{\tilde{s}\in\mathbb{R}^k} \frac{\gamma_a(\tilde{s},T)}{\tilde{r}_a(\tilde{s})}\,\mathrm{d}(\rho_s * \delta_s)(\tilde{s}),$$

where $*$ denotes convolution.

Consider the classification problem μ' derived from the original μ under an input transformation given by the joint distribution of $(f_A(X), A, Y)$, as discussed in Section 3.3, on which Id is the Bayes optimal predictor due to calibration of the f_a's. Then by Lemma B.5 applied on μ', the error rate on group a, denoted by $\mathrm{err}_a(\bar{h}_\rho)$, is

$$\mathrm{err}_a(\bar{h}_\rho) = \frac{1}{2}\int_{\Delta_k\times\mathcal{Y}} \|s-y\|_1\,\mathrm{d}\gamma'_a(s,y), \tag{9}$$

where $\gamma'_a \in \Gamma(r_a, q_a)$ equals to

$$\gamma'_a(s,y) = \int_{\mathrm{Id}^{-1}(s)} \mathcal{K}(s',y)\,\mathrm{d}r_a(s') = \mathcal{K}(s,y)\cdot r_a(s) = \int_{\mathbb{R}^k} \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\,\mathrm{d}(\rho_s * \delta_s)(\tilde{s})\cdot r_a(s),$$

whereby

$$\begin{aligned}
2\,\mathrm{err}_a(\bar{h}_\rho) &= \sum_{y\in\mathcal{Y}}\int_{\Delta_k}\int_{\mathbb{R}^k} \|s-y\|_1 \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\cdot(\rho_s*\delta_s)(\tilde{s}) r_a(s)\,\mathrm{d}\tilde{s}\,\mathrm{d}s \\
&\le \sum_{y\in\mathcal{Y}}\int_{\Delta_k}\int_{\mathbb{R}^k} \|\tilde{s}-y\|_1 \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\cdot(\rho_s*\delta_s)(\tilde{s}) r_a(s)\,\mathrm{d}\tilde{s}\,\mathrm{d}s \\
&\qquad + \sum_{y\in\mathcal{Y}}\int_{\Delta_k}\int_{\mathbb{R}^k} \|s-\tilde{s}\|_1 \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\cdot(\rho_s*\delta_s)(\tilde{s}) r_a(s)\,\mathrm{d}\tilde{s}\,\mathrm{d}s \\
&=: \sum_{y\in\mathcal{Y}}\int_{\mathbb{R}^k} \|\tilde{s}-y\|_1 \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\cdot\left(\int_{\Delta_k}(\rho_s*\delta_s)(\tilde{s}) r_a(s)\,\mathrm{d}s\right)\mathrm{d}\tilde{s} \\
&\qquad + \sum_{y\in\mathcal{Y}}\int_{\Delta_k}\int_{\mathbb{R}^k} \|n\|_1 \frac{\gamma_a(n-s,y)}{\tilde{r}_a(n-s)}\cdot(\rho_s*\delta_s)(n-s) r_a(s)\,\mathrm{d}n\,\mathrm{d}s \\
&= \sum_{y\in\mathcal{Y}}\int_{\mathbb{R}^k} \|\tilde{s}-y\|_1 \frac{\gamma_a(\tilde{s},y)}{\tilde{r}_a(\tilde{s})}\cdot\tilde{r}_a(\tilde{s})\,\mathrm{d}\tilde{s} + \int_{\Delta_k}\int_{\mathbb{R}^k}\|n\|_1\,\mathrm{d}\rho_s(n)\,\mathrm{d}r_a(s) \\
&= W_1(\tilde{r}_a, q_a) + \mathbb{E}_{N\sim\rho_s, s\sim r_a}[\|N\|_1],
\end{aligned}$$

where line 3 involves a change of variable $n \coloneqq \tilde{s} - s$. Then we have

$$\begin{aligned}
\mathrm{err}(\bar{h}_\rho) - \mathrm{err}^*_{\epsilon,f} &\le \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}(W_1(\tilde{r}_a,q_a) - W_1(r_a,q_a)) + \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}\mathbb{E}_{N\sim\rho_s,s\sim r_a}[\|N\|_1] \\
&\le \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}W_1(\tilde{r}_a,r_a) + \frac{1}{m}\sum_{a\in[m]}\frac{1}{2}\mathbb{E}_{N\sim\rho_s,s\sim r_a}[\|N\|_1].
\end{aligned}$$

Now, we upper bound the first term. Consider the coupling $\pi_a \in \Gamma(\tilde{r}_a, r_a)$ given by $\pi_a(\tilde{s}, s) = \rho_s(\tilde{s} - s)r_a(s)$, whereby

$$\begin{aligned} W_1(\tilde{r}_a, r_a) &= \inf_{\gamma \in \Gamma(\tilde{r}_a, r_a)} \int_{\mathbb{R}^k \times \Delta_k} \|\tilde{s} - s\|_1 \, \mathrm{d}\gamma(\tilde{s}, s) \\ &\leq \int_{\mathbb{R}^k \times \Delta_k} \|\tilde{s} - s\|_1 \, \mathrm{d}\pi_a(\tilde{s}, s) \\ &= \iint \|\tilde{s} - s\|_1 \rho_s(\tilde{s} - s) r_a(s) \, \mathrm{d}\tilde{s} \, \mathrm{d}s \\ &=: \iint \|(s + n) - s\|_1 \rho_s(n) r_a(s) \, \mathrm{d}n \, \mathrm{d}s \\ &= \mathbb{E}_{N \sim \rho_s, s \sim r_a}[\|N\|_1]. \end{aligned}$$

Substituting this into the result above, we obtain the upper bound.

On the other hand, the lower bound follows from Eq. (9), where

$$\mathrm{err}(\bar{h}_\rho) = \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma'_a(s, y) \geq \frac{1}{m} \sum_{a \in [m]} \frac{1}{2} W_1(r_a, q_a) = \mathrm{err}^*_{\epsilon, f}.$$

□

D Optimal Transport Between Simplex and Vertex Distributions

The $(k-1)$-dimensional probability simplex is defined for $k \geq 2$ by

$$\Delta_k \coloneqq \left\{ x \in \mathbb{R}^k : x \geq 0, \sum_{i=1}^k x_i = 1 \right\},$$

and its k vertices are $\{e_1, \cdots, e_k\}$. In this section, we study the optimal transportation problem between distributions supported on the simplex and its vertices under the ℓ_1 cost, given by $c(x, y) = \|x - y\|_1$.

By extending each Δ_k to infinity, we obtain a $(k-1)$-dimensional affine space of

$$\mathbb{D}^k \coloneqq \left\{ x \in \mathbb{R}^k : \sum_{i=1}^k x_i = 1 \right\} \supset \Delta_k.$$

Define vectors

$$v_{ij} \coloneqq e_j - e_i, \quad \forall i, j \in [k],$$

and note that for each $i \in [k]$, $\{v_{ij} : j \neq i\}$ forms a basis for $\mathbb{D}^k$. Also, observe the following identity for the ℓ_1 distance between a point on the simplex and a point on the vertex:

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

(this identity is central to some of the upcoming results).

A main result of this section is that when the transportation problem is semi-discrete, the deterministic (Monge) optimal transport exists, and is unique:

Theorem D.1. *Let p be a continuous probability measure on Δ_k, q a probability measure on $\{e_1, \cdots, e_k\}$, and $c(x, y) = \|x - y\|_1$. Then the optimal transport from p to q is a Monge plan, and is unique up to sets of measure zero w.r.t. p.*

Specifically, the optimal transport $\mathcal{T}^*_{p\to q}$ in Theorem D.1 is given by the c-transform of the Kantorovich potential from the Kantorovich-Rubinstein dual formulation of the transportation problem. In other words, it belongs to the following parameterized class of deterministic functions:

$$\mathcal{G}_k \coloneqq \left\{ x \mapsto e_{\arg\min_{i\in[k]}(\|x-e_i\|_1 - \psi_i)} : \psi \in \mathbb{R}^k \right\} \subset \{e_1, \cdots, e_k\}^{\Delta_k} \tag{11}$$

(break ties to the tied e_i with the largest index i).

This function class is therefore of particular interest to various analyses in this paper. For the generalization bounds in Section 4.2, we show that this function class has low complexity in terms of the Natarajan dimension (Definition D.12):

Theorem D.2. $d_{\mathrm{N}}(\mathcal{G}_k) = k - 1$.

In addition, note that as illustrated in Fig. 2, we can equivalently characterize each $g \in \mathcal{G}_k$ by the center point at which its k decision boundaries all intersect:

Proposition D.3. *Define the function class $\mathcal{G}'_k \subset \{e_1, \cdots, e_k\}^{\Delta_k}$ parameterized by $\mathbb{R}^k$ s.t. for each $g_z \in \mathcal{G}'_k$ with parameter $z \in \mathbb{R}^k$,*

$$g_z(x) = e_i \quad \text{if} \quad x_j - x_i \le z_j - z_i \iff x^\top v_{ij} \le z^\top v_{ij}, \quad \forall j \ne i \tag{12}$$

(when multiple e_i's are eligible, output the tied e_i with the largest index i).

Then $\forall g_\psi \in \mathcal{G}_k$, $g_\psi = g_z \in \mathcal{G}'_k$ by setting

$$z_i = \frac{1}{k} + \frac{1}{2}\left(\frac{1}{k}\sum_{j=1}^{k}\psi_j - \psi_i\right), \quad \forall i \in [k] \tag{13}$$

(the choice of $\sum_{i=1}^k z_i = 1$ s.t. $z \in \mathbb{D}^k$ was arbitrary, due to an extra degree of freedom because the support of g is contained in Δ_k).

Conversely, $\forall g_z \in \mathcal{G}'_k$, $g_z = g_\psi \in \mathcal{G}_k$ by setting

$$\psi_i = 2(z_1 - z_i) = 2z^\top v_{i1}, \quad \forall i \in [k] \tag{14}$$

(again, the choice of $\psi_1 = 0$ was arbitrary).

Proof. Let $g_\psi \in \mathcal{G}_k$, then for the $g_z \in \mathcal{G}'_k$ constructed in Eq. (13), by Eq. (10),

$$\begin{aligned}
g_z(x) = e_i \text{ is eligible} &\iff x_j - x_i \le z_j - z_i, && \forall j \ne i \\
&\iff x_j - x_i \le (\psi_i - \psi_j)/2, && \forall j \ne i \\
&\iff 2(x_j - x_i) \le \psi_i - \psi_j, && \forall j \ne i \\
&\iff 2(1 - x_i) - \psi_i \le 2(1 - x_j) - \psi_j, && \forall j \ne i \\
&\iff \|x - e_i\|_1 - \psi_i \le \|x - e_j\|_1 - \psi_j, && \forall j \ne i.
\end{aligned}$$

Conversely, let $g_z \in \mathcal{G}'_k$, then for the $g_\psi \in \mathcal{G}_k$ constructed in Eq. (14),

$$\begin{aligned}
g_\psi(x) = e_i \text{ is eligible} &\iff \|x - e_i\|_1 - \psi_i \le \|x - e_j\|_1 - \psi_j, && \forall j \ne i \\
&\iff 2(x_j - x_i) \le 2(z_1 - z_i) - 2(z_1 - z_j), && \forall j \ne i \\
&\iff x_j - x_i \le z_j - z_i, && \forall j \ne i.
\end{aligned}$$

□

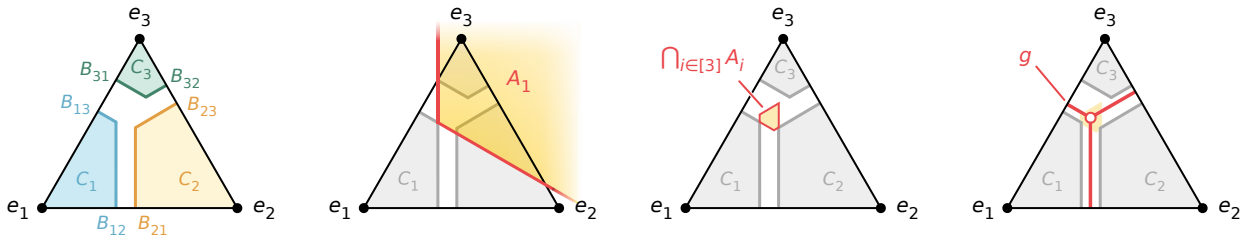

Figure 5: Illustration of the objects defined on Eqs. (16) and (17) for $k = 3$. See Fig. 6 for an example where the intersection is empty, when the underlying transport is not optimal.

We will often use this alternative characterization of $\mathcal{G}_k$.

The remaining proofs are deferred to Appendix D.2. Theorem D.1 is established via an analysis of the geometry of the simplex-vertex optimal transport, which we discuss in the next section.

D.1 Geometry of Optimal Transport

Let p be an arbitrary distribution supported on Δ_k, and q a (finite) distribution on $\mathcal{Y} \coloneqq \{e_1, \cdots, e_k\}$. We study the geometric properties of the optimal solution to the (Kantorovich) transportation problem between p, q under the ℓ_1 cost,

$$\sup_{\gamma \in \Gamma(p,q)} \int_{\Delta_k \times \mathcal{Y}} \|x - y\|_1 \, \mathrm{d}\gamma(x, y), \tag{15}$$

and note that the supremum can be attained because the supports are compact.

First, given a simplex-vertex transport $\gamma \in \Gamma(p, q)$, we define the following geometric objects:

$$\begin{aligned} B_{ij} &\coloneqq \min\Big\{ b \in \mathbb{R} : \gamma(\{x \in \Delta_k : x^\top v_{ij} \le b - 1\}, e_i) = q(e_i) \Big\} \cup \{0\}, \quad \text{and} \\ C_i &\coloneqq \bigcap_{j \neq i} \{x \in \Delta_k : x^\top v_{ij} \le B_{ij} - 1\}. \end{aligned} \tag{16}$$

For each $i \in [k]$, B_{ij} defines the (smallest offset of the) halfspace in the v_{ij} direction in which all points that are transported by γ to e_i are contained, and C_i is formed by the intersections of these halfspaces, also containing all points transported to e_i. See Figs. 3 and 5 for illustrations.

Now, if γ^* is an optimal transport of Eq. (15), then intuition tells us that in order to achieve minimum cost, the halfspaces along each direction should not overlap (i.e., $B_{ij} + B_{ji} \le 2$ for all $i \neq j$), and the C_i's should not intersect (except on a set of Lebesgue measure zero). We show that these intuitions regarding the geometry of γ^* are indeed valid, and they are implied by showing that the intersection of the following sets A_i is nonempty (see Fig. 5 for an illustration),

$$A_i \coloneqq \bigcap_{j \neq i} \{x \in \mathbb{D}^k : x^\top v_{ij} \ge B_{ij} - 1\}. \tag{17}$$

Proposition D.4. *If γ^* is a minimizer of Eq. (15), then $\bigcap_{i\in[k]} A_i \neq \emptyset$.*

The proof is deferred to Appendix D.2. Note that $\bigcap_{i\in[k]} A_i$ is exactly the set considered on Line 7 of Algorithm 2, and Proposition D.4 says that if γ^* is an optimal transport, then a point $z \in \bigcap_{i\in[k]} A_i$ exists. The significance of this point is that, the function $g_z \in \mathcal{G}_k$ with parameter $z \in \mathbb{D}^k$ agrees with the transport $\mathcal{T}_{p\to q}$ associated with γ^* only except for points that lie on the boundaries (which have Lebesgue measure zero):

Lemma D.5. *Let p, q be probability measures on Δ_k and $\{e_1, \cdots, e_k\}$, respectively. If $\gamma^* \in \Gamma(p, q)$ is a minimizer of Eq.* (15)*, then $\exists \mathcal{T} \in \mathcal{G}_k$ with parameters $z \in \mathbb{D}^k$ satisfying*

$$\gamma(x, \mathcal{T}(x)) = p(x), \quad \forall x \in \operatorname{supp}(p) \setminus \bigcup_{i \neq j} \{x \in \mathbb{D}^k : x^\top v_{ij} = z^\top v_{ij}\}.$$

This result underlies many discussions throughout our presentation: (1) the construction used in its proof led to Lines 6–9 of Algorithm 2 for extracting post-processing functions from the empirical optimal transports, (2) it embodies the argument used in the proof of Theorem 4.4 regarding the disagreements between the extracted functions and the empirical transports, and (3) the existence part of Theorem D.1 is a direct consequence, since the set on which disagreements may occur always has measure zero when p is continuous.

Proof. Let $z \in \bigcap_{i \in [k]} A_i$, which exists due to Proposition D.4. Then let $\mathcal{T} \in \mathcal{G}_k$ with parameter z, which we show agrees with γ^* on all $x \in \operatorname{supp}(p) \setminus \bigcup_{i \neq j} \{x \in \mathbb{D}^k : x^\top v_{ij} = z^\top v_{ij}\}$: suppose $\mathcal{T}(x) = e_i$, then $\mathcal{T}(x) = e_i \iff x^\top v_{ij} \leq z^\top v_{ij}$ by construction. Furthermore, by the definition of A_i in Eq. (17) of γ^*, $x^\top v_{ij} < z^\top v_{ij} \leq B_{ij} - 1$ for all $j \neq i$, so we must have that $\gamma^*(x, e_j) = 0$, $\forall j \neq i \implies \gamma^*(x, e_i) = p(x)$. Otherwise, it would contradict the definition of B_{ij} in Eq. (16). □

D.2 Omitted Proofs from Section D

Proof of Theorem D.1. For existence, Lemma D.5 provides a $\mathcal{T} \in \mathcal{G}_k$ that agrees with the optimal transport almost everywhere, since the set of points lying on the boundaries has measure zero w.r.t. p by continuity.

Next, we prove uniqueness. Let $\gamma, \gamma' \in \Gamma(p, q)$ be two optimal transports, and $\mathcal{T}, \mathcal{T}' \in \mathcal{G}_k$ mappings provided by Lemma D.5 that agree with γ, γ' a.e. We will show that $\mathcal{T} = \mathcal{T}'$ a.e., and so is $\gamma = \gamma'$.

Denote the parameter (i.e. center of the decision boundaries) of $\mathcal{T}$ (analogously for $\mathcal{T}'$) by $z \in \mathbb{D}^k$, the decision boundaries by $B_{ij} \coloneqq z^\top v_{ij} + 1$, and the decision regions by $C_i \coloneqq \bigcap_{j \neq i} \{x \in \Delta_k : x^\top v_{ij} \leq B_{ij} - 1\}$. By definition of $\mathcal{G}_k$, $\Delta_k = \bigsqcup_{i=1}^k C_i$, and for all $x \in \Delta_k$, $\mathcal{T}(x) = e_i \iff x \in C_i$ almost surely, therefore, $\gamma(C_i, e_i) = q(e_i)$ because $\mathcal{T}$ agrees with the transport γ a.e.

Define the difference in the boundaries between $\mathcal{T}$ and $\mathcal{T}'$ by $d_{ij} \coloneqq B'_{ij} - B_{ij}$, and note that

$$d_{ij} = d_{nj} - d_{ni} \quad \text{with} \quad d_{\ell\ell} \coloneqq 0, \quad \forall i, j, n, \ell \in [k], \tag{18}$$

which follows from the observation that

$$B_{ij} = z^\top (v_{nj} - v_{ni}) + 1 = B_{nj} - B_{ni} + 1 \quad \text{with} \quad B_{\ell\ell} \coloneqq 0, \quad \forall i, j, n, \ell \in [k].$$

Construct a directed graph of k nodes where (i, j) is an edge iff $d_{ij} > 0$. Note that this graph is acyclic: first, it cannot contain cycles of length 2, otherwise, $(i, j), (j, i) \in E \implies d_{ij} + d_{ji} > 0$ contradicts the fact that $d_{ij} + d_{ji} = 0$ by definition; next, consider the shortest cycle, and let $(i, j), (n, i)$, $j \neq n$ denote two edges contained in it. It follows that $d_{ij}, d_{ni} > 0$, and $d_{nj} \leq 0$, or it is not the shortest cycle. Then by Eq. (18), $0 < d_{ij} = d_{nj} - d_{ni} < 0$, which is a contradiction.

Now, we show by strong induction on the reverse topological order of the graph nodes that for all $i \in [n]$, $p(C_i \oplus C'_i) = 0$ where $\oplus$ denotes the symmetric difference of the sets. For the base case, let i denote a sink node in the graph, then we have that $d_{ij} \leq 0$ for all j, meaning that $C'_i \subseteq C_i$. Then $q(e_i) = \gamma'(C'_i, e_i) = p(C'_i) \leq p(C_i) = \gamma(C_i, e_i) = q(e_i)$. If the inequality is strict, then it is a contradiction; otherwise, combining the equality with $\mathcal{T}(x) = e_i \iff x \in C_i$ a.s. (and $\mathcal{T}'$

analogously) implies $p(C_i \oplus C_i') = p(C_i \setminus C_i') = 0$. For the inductive case, let i denote a node, and $J \subseteq [n] \setminus \{i\}$ the set of nodes directed to from i, then by construction $\bigsqcup_{j \in J \cup \{i\}} C_j' \subseteq \bigsqcup_{j \in J \cup \{i\}} C_j$. Let $F_i \coloneqq C_i \cap C_i'$, and note that for all $x \in C_i' \setminus F_i$, $\mathcal{T}'(x) = e_i$ and $\mathcal{T}(x) \in \{e_j : j \in J \setminus \{i\}\}$. Therefore, $C_i' \setminus F_i \in \bigcup_{j \in J}(C_j \oplus C_j')$, and by the inductive hypothesis, $p(C_i' \setminus F_i) \leq 0$. It then follows that $p(C_i') \leq p(C_i)$, and subsequently $p(C_i \oplus C_i') = p(C_i \setminus C_i') = 0$ by the same arguments used in the base case.

Therefore, $p(\{x : \mathcal{T}(x) \neq \mathcal{T}'(x)\}) \leq \sum_{i=1}^{k} p(C_i \oplus C_i') = 0$, so $\mathcal{T} = \mathcal{T}'$ a.e. □

The proof of Proposition D.4 needs the following technical result, which at a high-level states that if a collection of v_{ij}-aligned convex sets do not intersect, then they cannot cover the entire space:

Proposition D.6. *Let $B \in \mathbb{R}^{k \times k}$ arbitrary, and define $S_i \coloneqq \bigcap_{j \neq i}\{x \in \mathbb{D}^k : x^\top v_{ij} \leq B_{ij} - 1\}$ for each $i \in [k]$, then $\bigcap_{i \in [k]} S_i = \emptyset \implies \bigcup_{i \in [k]} S_i \neq \mathbb{D}^k$.*

While this could be proved with elementary arguments, for clarity, we use known results from algebraic topology in the final steps of our proof. The tools and concepts that we use include homotopy equivalence and homology groups (we omit the definition for the latter, but refer readers to (Spanier, 1981) for a textbook). The definitions are provided below for completeness; readers may skip to the main proof.

Definition D.7 (Homotopy)**.** Let $\mathcal{X}, \mathcal{Y}$ be topological spaces, and $f, g : \mathcal{X} \to \mathcal{Y}$ two continuous functions. A homotopy between f and g is a continuous function $h : \mathcal{X} \times [0,1] \to \mathcal{Y}$, such that $h(x,0) = f(x)$ and $h(x,1) = g(x)$ for all $x \in \mathcal{X}$. We say f, g are homotopic if there exists a homotopy between them.

Definition D.8 (Homotopy Equivalence)**.** Let $\mathcal{X}, \mathcal{Y}$ be topological spaces. If there exist continuous maps $f : \mathcal{X} \to \mathcal{Y}$ and $g : \mathcal{Y} \to \mathcal{X}$ such that $g \circ f$ is homotopic to the identity map $\mathrm{Id}_{\mathcal{X}}$ on $\mathcal{X}$, and $f \circ g$ is homotopic to $\mathrm{Id}_{\mathcal{Y}}$, then $\mathcal{X}$ and $\mathcal{Y}$ are homotopy equivalent, denoted by $\mathcal{X} \cong \mathcal{Y}$.

Fact D.9 (Homology)**.** *(See (Spanier, 1981) for a textbook).*

1. *The homology groups of $\mathbb{R}^d$, denoted by $H_n(\mathbb{R}^d)$ for $n \in \{0, 1, 2, \cdots\}$, are*

$$H_n(\mathbb{R}^d) = \begin{cases} \mathbb{Z} & \text{if } n = 0 \\ \{0\} & \text{else.} \end{cases}$$

2. *The homology groups of the d-dimensional simplex, Δ_{d+1}, are*

$$H_n(\Delta_{d+1}) = \begin{cases} \mathbb{Z} & \text{if } n = 0 \\ \{0\} & \text{else.} \end{cases}$$

3. *The homology groups of the d-dimensional simplex without its interior, $\partial\Delta_{d+1}$, are*

$$H_n(\partial\Delta_{d+1}) = \begin{cases} \mathbb{Z} & \text{if } n = 0 \text{ or } d-1 \\ \{0\} & \text{else.} \end{cases}$$

4. *If $\mathcal{X} \cong \mathcal{Y}$, then $H_n(\mathcal{X}) = H_n(\mathcal{Y})$ for all n.*

Clearly, the affine space $\mathbb{D}^k \cong \mathbb{R}^{k-1}$ via a rotation and a translation. We also cite the Nerve theorem (Bauer et al., 2022, Theorem 3.1):

Theorem D.10 (Nerve)**.** *Let $\mathcal{S} = \{S_1, \cdots, S_n\}$ be a finite collection of sets, and define its* nerve *by*

$$\mathrm{Nrv}(\mathcal{S}) = \left\{ J \subseteq [n] : \bigcap_{i \in J} S_i \neq \emptyset \right\}.$$

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

Proof of Proposition D.6. 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 on the dimensionality k. For the base case of $k = 2$, observe that $S_1 \cup S_2 = \{x : x^\top v_{12} \leq B_{12} - 1 \text{ or } x^\top v_{12} \geq 1 - B_{21}\}$, 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 $(1 - B_{12}/2, B_{12}/2) \in S_1 \cap S_2$, thereby the intersection is nonempty.

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

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

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

$$S'_{J,i} \coloneqq \bigcap_{j\in J, j\neq i} \{x \in \mathbb{D}^k : x^\top v_{ij} \leq B_{ij} - 1\},$$

(we will drop the subscript J as the discussions below will focus on a single 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, $\mathbb{D}^k = \bigcup_{i\in[k]} S_i \subset \bigcup_{i\in[k]} S'_i$, and we argue that $\bigcup_{i\in J} S'_i = \mathbb{D}^k$. Suppose not, then let $z \notin \bigcup_{i\in J} S'_i$, and consider the line

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

First, 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

$$v_{ij}^\top \sum_{n\notin J, m\in J} v_{nm} = \sum_{n\notin J, m\in J} (e_j^\top e_m - e_j^\top e_n - e_i^\top e_m + e_i^\top e_n) = 1 - 0 - 1 + 0 = 0.$$

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

$$v_{ij}^\top \sum_{n\notin J, m\in J} v_{nm} = \sum_{n\notin J, m\in J} (e_j^\top e_m - e_j^\top e_n - e_i^\top e_m + e_i^\top e_n) = 1 - 0 - 0 + 1 = 2;$$

so points on L with sufficiently large α's are not contained in $\bigcup_{i\notin J} S'_i$, contradicting the assumption that $\bigcup_{i\in[k]} S'_i = \mathbb{D}^k$.

Back to proving that $\bigcap_{i\in J} S_i \neq \emptyset$ for any $J \subset [k]$ with $|J| \leq k - 1$. Since $\bigcup_{i\in J} S'_i = \mathbb{D}^k$, by applying the inductive hypothesis to a $|J|$-dimensional instance derived from $\{S'_i : i \in J\}$ by removing the axes $\{e_i : i \notin J\}$, we get $\exists z' \in \bigcap_{i\in J} S'_i$. Using similar arguments above, it can be shown that the line $L' \coloneqq \{z' + \alpha \sum_{i\notin J, j\in J} v_{ij} : \alpha \in \mathbb{R}\}$ is entirely contained in $\bigcap_{i\in J} S'_i$ and partially 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 the intersection is nonempty.

We have thus established that any intersection of the strict subset of $\{S_i\}_{i\in[k]}$ is nonempty, and we will conclude with the Nerve theorem. We have $\forall J \subset [k]$, $1 \leq |J| \leq k - 1$, $J \in$

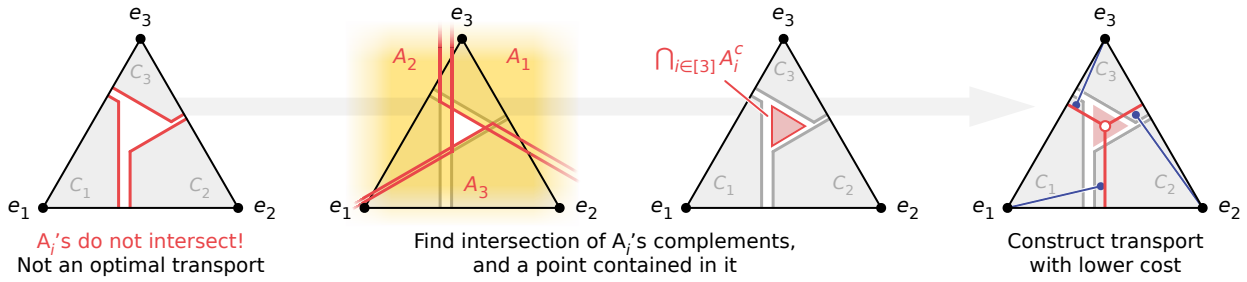

Figure 6: Illustration of the construction in the proof of Proposition D.4 for $k = 3$.

$\mathrm{Nrv}(\{S_1, \cdots, S_k\})$. Because we assumed in the beginning that $\bigcup_{i\in[k]} S_i = \mathbb{D}^k$, it must follow that $[k] \in \mathrm{Nrv}(\{S_1, \cdots, S_k\})$ as well. Otherwise, the nerve is a $(k-1)$-dimensional simplex (each n-face is represented by its $n-1$ vertices) without its interior (represented by $[k]$), whose homology differs from that of $\mathbb{D}^k$, then $\bigcup_{i\in[k]} S_i \cong \mathrm{Nrv}(\{S_1, \cdots, S_k\}) \ncong \mathbb{D}^k$ by Theorem D.10, which contradicts our assumption that $\bigcup_{i\in[k]} S_i = \mathbb{D}^k$. Hence the nerve contains $[k]$, meaning $\bigcap_{i\in[k]} S_i \neq \emptyset$. □

Proof of Proposition D.4. Recall the definitions of the objects B_{ij}, C_i and A_i in Eqs. (16) and (17) of γ^*. Suppose $\bigcap_{i\in[k]} A_i = \emptyset$, then $\exists z \in \bigcap_{i\in[k]}(\mathbb{D}^k \setminus A_i)$ by Proposition D.6. It then follows by definition that $\forall i \in [k]$, $\exists j \neq i$ s.t. $z^\top v_{ij} < B_{ij} - 1$. Let $u : [k] \to [k]$ denote a mapping s.t. the pairs $(i, u(i))$ satisfy this relation for all $i \in [k]$; note that there exists a nonempty $J \subseteq [m]$ s.t. the undirected edges $\{(i, u(i)) : i \in J\}$ form a cycle because $u(i) \neq i$. Also, there exist $m > 0$ and measurable sets $F_i \subset \{x : x^\top v_{iu(i)} > z^\top v_{iu(i)}\} \subset C_i$ s.t. $\gamma^*(F_i, e_i) \coloneqq m_i \geq m$.

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

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

has a lower transportation cost than γ^* (see Fig. 6 for an illustration):

$$\begin{aligned} \int_{\Delta_k \times \mathcal{Y}} \|x - y\|_1 \,\mathrm{d}(\gamma^* - \gamma')(x, y) &= \sum_{i\in J} \frac{m}{m_i} \int_{F_i} \left(\|x - e_i\|_1 - \|x - e_{u(i)}\|_1\right) \mathrm{d}\gamma^*(x, e_i) \\ &= \sum_{i\in J} \frac{2m}{m_i} \int_{F_i} x^\top v_{iu(i)} \,\mathrm{d}\gamma^*(x, e_i) \\ &> \sum_{i\in J} \frac{2m}{m_i} \int_{F_i} z^\top v_{iu(i)} \,\mathrm{d}\gamma^*(x, e_i) \\ &= 2m \sum_{i\in J} z^\top v_{iu(i)} \\ &= 2m \sum_{i\in J} (z_{u(i)} - z_i) = 0, \end{aligned}$$

where line 2 follows from Eq. (10). □

Finally, we consider the complexity of the function class $\mathcal{G}_k$ defined in Eq. (11). First, recall the (multiclass-generalized) definition of shattering, based on which the Natarajan dimension is defined (Shalev-Shwartz and Ben-David, 2014, Definitions 29.1 and 29.2):

Definition D.11 (Shattering)**.** Let $\mathcal{H} \subset \{1, \cdots, k\}^{\mathcal{X}}$ be a class of multiclass functions. $\mathcal{H}$ is said to shatter a set $S \subset \mathcal{X}$ if there exist $f_0, f_1 : S \to \{1, \cdots, k\}$ satisfying $f_0(x) \neq f_1(x)$ for all $x \in S$, such that $\forall S_0, S_1 \subset S$ that partition S, $\exists h \in \mathcal{H}$,

$$\begin{aligned} h(x) &= f_0(x), \quad \forall x \in S_0, \quad \text{and} \\ h(x) &= f_1(x), \quad \forall x \in S_1. \end{aligned}$$

Definition D.12 (Natarajan Dimension)**.** The Natarajan dimension of a class of multiclass functions $\mathcal{H} \subset \{1, \cdots, k\}^{\mathcal{X}}$, denoted by $d_{\mathrm{N}}(\mathcal{H})$, is the largest number n s.t. $\exists S \subset \mathcal{X}$ of cardinality n that is shattered by $\mathcal{H}$.

Proof of Theorem D.2. We associate e_i with the label i, $\forall i \in \{1, \cdots, k\}$. We first show that $d_{\mathrm{N}}(\mathcal{G}_k) \geq k - 1$ by constructing a set of cardinality $k - 1$ that is shattered by $\mathcal{G}_k$, then show that $d_{\mathrm{N}}(\mathcal{G}_k) < k$ by contradiction.

Lower Bound. Consider the set $S = \{e_1, e_2, \cdots, e_{k-1}\}$ and let $f_0(e_j) = j$ and $f_1(e_j) = k$ for all $j \in [k-1]$, which satisfy $f_0 \neq f_1$ on all $x \in S$. Let $S_0 \sqcup S_1 = S$ be arbitrary, and define

$$\iota(j) \coloneqq \begin{cases} \mathbb{1}(e_j \in S_1) & \text{if } j \in [k-1] \\ 0 & \text{if } j = k. \end{cases}$$

Consider $g_z \in \mathcal{G}_k$ with parameters

$$z = \frac{1}{k} \cdot \mathbf{1}_k - \sum_{j=1}^{k-1} \iota(j) \sum_{\ell \neq j} v_{j\ell},$$

where boldface $\mathbf{1}_k \in \mathbb{R}^k$ denotes the vector of all ones. Observe that

$$\begin{aligned} z^\top v_{nm} &= -\sum_{j=1}^{k-1} \iota(j) \sum_{\ell \neq j} v_{j\ell}^\top v_{nm} \\ &= -\sum_{j=1}^{k-1} \iota(j) \sum_{\ell \neq j} (e_\ell^\top e_m - e_\ell^\top e_n - e_j^\top e_m + e_j^\top e_n) \\ &= \sum_{j=1}^{k-1} \iota(j)(\mathbb{1}(n \neq j) - \mathbb{1}(m \neq j)) + (k-1) \sum_{j=1}^{k-1} \iota(j)(e_j^\top e_m - e_j^\top e_n) \\ &= (k-1)(\iota(m) - \iota(n)) + (\iota(m) - \iota(n)) \\ &= k(\iota(m) - \iota(n)). \end{aligned}$$

Recall from Eq. (12) that for all $i, n \in [k]$,

$$g_z(e_i) = e_n \text{ is eligible } \iff e_i^\top v_{nm} \leq z^\top v_{nm}, \quad \forall m \neq n;$$

so in our case, it follows that for all $i \in [k-1]$ and $j \neq i$,

$$g_z(e_i) = e_i \text{ is eligible } \iff -1 \leq k(\iota(m) - \iota(i)), \quad \forall m \neq i,$$

and

$$g_z(e_i) = e_j \text{ is eligible } \iff 1 \leq k(\iota(i) - \iota(j)) \quad \text{and} \quad 0 \leq k(\iota(m) - \iota(j)), \quad \forall m \neq i$$

(also, recall $\iota(k) \coloneqq 0$).

Observe that for any $i \in [k-1]$, if $\iota(i) = 0$, then $g_z(e_i) = e_j$ is ineligible for all $j \neq i$, then we must have $g_z(e_i) = e_i$. Otherwise, if $\iota(i) = 1$, then e_k is always eligible, so $g_z(e_i) = e_k$ due to the tie-breaking rule. Therefore, g_z is a witness function, and we conclude that $\mathcal{G}_k$ shatters S.

Upper Bound. Let $S = (x_1, \cdots, x_k)$ be given, along with $f_0, f_1 : S \to [k]$ satisfying $f_0(x) \neq f_1(x)$ for all $x \in S$. Suppose $\mathcal{G}_k$ shatters S. Let $g_z \in \mathcal{G}_k$ denote a witness function for the partitioning of $S_0 = S$ and $S_1 = \emptyset$, and $g_{z'} \in \mathcal{G}_k$ that for the partitioning of $S'_0 = \emptyset$ and $S'_1 = S$.

We will reuse an argument from an earlier proof. Denote the decision boundaries of g_z (analogously for $g_{z'}$) by $B_{ij} \coloneqq z^\top v_{ij} + 1$, and the decision regions by $C_i \coloneqq \bigcap_{j \neq i} \{x \in \Delta_k : x^\top v_{ij} \leq B_{ij} - 1\}$. By definition of $\mathcal{G}_k$, $x \in C_i \implies g_z(x) = e_i$ is eligible. Then, define the difference in the boundaries between g_z and $g_{z'}$ by $d_{ij} \coloneqq B'_{ij} - B_{ij}$. Construct a directed graph of k nodes where (i, j) is an edge iff $d_{ij} > 0$, which is acyclic as shown in the proof of Theorem D.1.

First, consider the case where $\exists x, x' \in S$ s.t. $\{f_0(x), f_1(x)\} = \{f_0(x'), f_1(x')\}$. W.l.o.g., assume $i \coloneqq f_0(x_1) = f_1(x_2)$ and $j \coloneqq f_1(x_2) = f_0(x_1)$, then we have

$$g_z(x_1) = e_i,\ g_{z'}(x_1) = e_j \implies d_{ji} > 0,$$
$$g_z(x_2) = e_j,\ g_{z'}(x_2) = e_i \implies d_{ij} > 0$$

(after taking into account of the tie-breaking rule), however, this would imply a cycle in the graph, which is a contradiction.

Next, if $\{f_0(x), f_1(x)\}$ differs for all $x \in S$, then we may assume w.l.o.g. $f_0(x_i) = e_i$ and $f_1(x_i) = e_{i+1}$ for all $i \in [k]$ (where the index of $k+1$ means 1). Then

$$g_z(x_i) = e_i,\ \ g_{z'}(x_i) = e_{i+1} \implies d_{i+1,i} > 0, \quad \forall i \in [k];$$

again, this would imply a cycle in the graph, hence a contradiction. Therefore, we conclude that $\mathcal{G}_k$ cannot shatter any $S \subset \Delta_k$ of cardinality k. □

In addition, on data distributions that satisfy $X = Y$, $X \in \Delta_k$, applying the ℓ_1 loss of $\ell(\hat{y}, y) \coloneqq \|\hat{y} - y\|_1$ to $\mathcal{G}_k$ yields a function class with pseudo-dimension of $k - 1$:

Definition D.13 (Pseudo-Shattering)**.** Let $\mathcal{H} \subset [0,1]^{\mathcal{X}}$ be a class of real-valued functions. $\mathcal{H}$ is said to pseudo-shatter a set $\{x_1, \cdots, x_n\} \subset \mathcal{X}$ if $\exists r_1, \cdots, r_n \in \mathbb{R}$ such that $\forall y_1, \cdots, y_n \in \{0, 1\}$, $\exists h \in \mathcal{H}$ satisfying $\mathbb{1}(h(x_i) \geq r_i) = y_i$ for all $i \in [n]$.

Definition D.14 (Pseudo-Dimension)**.** The pseudo-dimension of a class of real-valued functions $\mathcal{H} \subset [0,1]^{\mathcal{X}}$, denoted by $d_{\mathrm{P}}(\mathcal{H})$, is the largest number n s.t. $\exists S \subset \mathcal{X}$ of cardinality n that is pseudo-shattered by $\mathcal{H}$.

Theorem D.15. *Define* $\mathcal{H}_k \coloneqq \{x \mapsto \|g(x) - x\|_1 : g \in \mathcal{G}_k\}$*, then* $d_{\mathrm{P}}(\mathcal{H}_k) = k - 1$.

Proof. The proof shares the same arguments as that of Theorem D.2. We will only show the upper bound, and remark that the lower bound can be established using a similar construction of that in Theorem D.2.

Let $x_1, \cdots, x_k$ be given, and suppose there exists thresholds $r_1, \cdots, r_k$ s.t. $\mathcal{H}_k$ shatters the set of points. It follows that $\exists g_z, g_{z'} \in \mathcal{G}_k$ s.t. $\|g_{z'}(x_i) - x_i\|_1 < r_i \leq \|g_z(x_i) - x_i\|_1$ for all i, which means that $g_z(x_i) \neq g_{z'}(x_i)$. But by the arguments in the proof of the upper bound of Theorem D.2, such $(g_z, g_{z'})$ pair does not exists, contradicting the shattering assumption. □

Finally, for all $i \in \{1, \cdots, k\}$, define the label-i-versus-all binarization of $\mathcal{G}_k$ by

$$\mathcal{G}_{k,i} \coloneqq \{x \mapsto \mathbb{1}(g(x) = e_i) : g \in \mathcal{G}_k\} \subset \{0, 1\}^{\Delta_k}. \tag{19}$$

Clearly, its *VC dimension*, the binary class reduction of the Natarajan dimension (Definition D.12), denoted by $d_{\mathrm{VC}}(\mathcal{G}_{k,i})$, is upper bounded by $d_{\mathrm{N}}(\mathcal{G}_k)$:

Corollary D.16. $d_{\mathrm{VC}}(\mathcal{G}_{k,i}) \leq d_{\mathrm{N}}(\mathcal{G}_k) = k - 1$.

E Experiment Details

E.1 Datasets and Tasks

We provide descriptions of each dataset and the tasks they entail. Statistics of the datasets are included in Tables 3 to 5, where the numbers indicate the proportion (%) of examples belonging to the classes in each group.

Adult (Kohavi, 1996)**.** This binary classification task decides whether the annual income of an individual is over or below \$50k per year ($|\mathcal{Y}| = 2$) given attributes including gender, race, age, education level, etc. The data are collected from the 1994 US Census. We consider gender to be the sensitive attribute ($|\mathcal{A}| = 2$ from the dataset).

The tabular dataset is pre-processed by normalizing numerical attributes and converting categorical attributes into one-hot representation. The training set of 30,162 examples is split in half for training the predictor and performing post-processing.

Communities & Crime (Redmond and Baveja, 2002)**.** The Communities & Crime tabular dataset contains the socioeconomic and crime data of communities in 46 US 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, 1990 LEMAS survey, and 1995 FBI Uniform Crime Reporting program. We bin the rate of violent crime into five classes ($|\mathcal{Y}| = 5$), and we treat race as the sensitive attribute by the presence of minorities ($|\mathcal{A}| = 4$): 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 is the same as that on Adult. Because of the small dataset size of 1,994, results are averaged with 10-fold cross validation (using the folds defined and provided in the dataset).

BiasBios (De-Arteaga et al., 2019)**.** The task is to determine the occupation ($|\mathcal{Y}| = 28$) of female and male individuals ($|\mathcal{A}| = 2$) by their raw text biographies. The data are mined from the Common Crawl corpus. In this dataset, gender is the sensitive attribute, and is observed to correlate with certain occupations such as software engineer and nurse.

We use the version of BiasBios compiled and hosted by Ravfogel et al. (2020), which contains 255,710 training and 98,344 test examples. We split the training set by 95%/5% for training the predictor and performing post-processing, respectively.

This experiment is of particular interest because of the increasing popularity of large language models and the fairness concerns regarding their usage. In particular, the uncurated corpora (e.g., crawled from the internet) on which the language models are pre-trained may contain historical social bias, and empirical investigations have shown that such bias could be propagated and amplified in downstream applications (Bolukbasi et al., 2016; Zhao et al., 2018; Abid et al., 2021).

E.2 Experiment Setup

On each dataset, we first train attribute-aware predictors via minimizing the MSE w.r.t. the one-hot class labels without constraints (Eq. (7)), then post-process the predictors using Algorithm 2, and also the smoothing procedure in Section 4.3 (Appendix E.3). Note that our post-processing

Table 2: Group-balanced error rate (Eq. (4)) and DP fairness (Δ_{DP}; Definition 2.1) of classifiers post-processed using Algorithm 2 and the smoothing procedure of Section 4.3, for strict DP ($\epsilon = 0$). The number of groups $\mathcal{A}$ and classes $\mathcal{Y}$ on each task are included.

Dataset	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{Y}\rvert$	Method	err	Δ_{DP}
Adult	2	2	Not post-processed	0.1365	0.1556
			Algorithm 2	0.1566	0.00009
			Alg. 2 w/ smoothing	0.1566	0.00005
Communities & Crime	4	5	Not post-processed	0.3264	0.5321
			Algorithm 2	0.3928	0.1406
			Alg. 2 w/ smoothing	0.3836	0.1168
BiasBios	2	28	Not post-processed	0.1414	0.2865
			Algorithm 2	0.1949	0.1352
			Alg. 2 w/ smoothing	0.2091	0.1188

procedure is also applicable to models trained with other losses, e.g., cross-entropy loss, but the performance may be suboptimal without model calibration as discussed in Section 3.3.

We use linear models on the Adult and Communities & Crime tabular datasets (learned via OLS), and a neural language model fine-tuned from the BERT base (uncased) checkpoint with 110 million parameters on the BiasBios text dataset (Devlin et al., 2019). Recall that Algorithm 2 requires the outputs of the predictor to be contained in the simplex Δ_k; we do not enforce this constraint during training, and instead project the outputs to the simplex during post-processing and inference. For simplex projection, we use the implementation by Blondel et al. (2014).

BERT Hyperparameters. To fine-tune BERT base (uncased) pre-trained language model on BiasBios, we add a new linear layer on the output embedding of the `[CLS]` token, and train the model end-to-end for three epochs using the AdamW optimizer. We set the batch size to 32, learning rate to 2e-5, use a linearly decaying learning rate schedule with a warmup ratio of 0.1, L^2 weight decay rate of 0.01, and clip the norm of the gradients to 1. The PyTorch implementation is included in our code,[2] which uses the Hugging Face Transformers library (Wolf et al., 2020).

E.3 Results with Smoothing Procedure

Since it may be possible that the continuous assumption on the r_a's required by Algorithm 2 is not satisfied by our pre-trained predictors, we explore whether further performance gains could be obtained from applying the remedy described in Section 4.3 of smoothing the distributions by a continuous noise.

For the continuous noise, we use a distribution based on the Dirichlet distribution, with pointwise scaling to ensure roughly equal variance on each input point (we refer readers to our code[2] for the implementation). Using the Dirichlet distribution ensures that the perturbed points always stay inside the simplex, hence we can apply the procedure used in Algorithm 2 to find the optimal transport from the smoothed $\tilde{r}_a$ to each $\hat{q}_a$, the latter of which are estimated from calling $\mathrm{OPT}(\hat{r}_1, \cdots, \hat{r}_m, \epsilon)$ (note that it is not called on the smoothed versions of the empirical distributions, which would be denoted by $\tilde{\hat{r}}_a$).

Each example is perturbed 10 times for post-processing, and 1000 times during inference. The results under $\epsilon = 0$ are presented in Table 2, where we observe that smoothing achieves significant further improvements to Δ_{DP}.

E.4 Finding Feasible Point in Intersection of Halfspaces

Line 7 of Algorithm 2 involves finding a feasible point in the intersection of halfspaces, which can be obtained with the following linear program:

$$\min_{z \in \mathbb{R}^k} 0 \quad \text{s.t.} \quad z^\top v_{ij} \leq B_{ij} - 1, \quad \forall i, j \in [k],\, i \neq j.$$

As illustrated in Fig. 3, the point z that is returned determines the center of the boundaries of the extracted transport maps $\mathcal{T}_a \in \mathcal{G}_k$. Because of the machine learning folklore that classifiers with larger margin enjoy better generalization properties, we instead use the follow quadratic program (a least-squares problem) that maximizes the margins in our experiments for point-finding:

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

Although comparisons are not included, our preliminary experiments showed that using the QP for point-finding led to better post-processing performance with Algorithm 2 than using the LP, both in terms of the error rate and Δ_{DP} during inference.

Table 3: Dataset statistics of Adult (%).

Class \ Group	Female	Male
$\leq$ 50k	89.07	69.62
$>$ 50k	10.93	30.38
Count	14423	22732

Table 4: Dataset statistics of Communities & Crime (%).

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

Table 5: Dataset statistics of BiasBios (%).

Class \ Group	Female	Male
Accountant	1.14	1.69
Architect	1.32	3.65
Attorney	6.86	9.52
Chiropractor	0.38	0.90
Comedian	0.33	1.04
Composer	0.50	2.22
Dentist	2.83	4.41
Dietitian	2.03	0.14
DJ	0.12	0.60
Filmmaker	1.27	2.22
Interior Designer	0.65	0.13
Journalist	5.42	4.77
Model	3.41	0.61
Nurse	9.47	0.82
Painter	1.95	1.98
Paralegal	0.82	0.13
Pastor	0.33	0.91
Personal Trainer	0.36	0.37
Photographer	4.77	7.40
Physician	10.75	8.98
Poet	1.89	1.69
Professor	29.26	30.67
Psychologist	6.25	3.27
Rapper	0.07	0.60
Software Engineer	0.60	2.75
Surgeon	1.08	5.35
Teacher	5.36	3.04
Yoga Teacher	0.77	0.12
Count	182102	211321