Fair and Optimal Classification via Post-Processing

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

Abstract

To mitigate the bias exhibited by machine learning models, fairness criteria can be integrated into the training process to ensure fair treatment across all demographics, but it often comes at the expense of model performance. Understanding such tradeoffs, therefore, underlies the design of fair algorithms. To this end, this paper provides a complete characterization of the inherent tradeoff of *demographic parity* on classification problems, under the most general multi-group, multi-class, and noisy setting. Specifically, we show that the minimum error rate achievable by randomized and attribute-aware fair classifiers is given by the optimal value of a Wasserstein-barycenter problem. On the practical side, our findings lead to a simple post-processing algorithm that derives fair classifiers from score functions, which yields the optimal fair classifier when the score is Bayes optimal. We provide suboptimality analysis and sample complexity for our algorithm, and demonstrate its effectiveness on benchmark datasets.

1. Introduction

Machine learning models trained on biased data have been found to perpetuate and even amplify the bias against historically underrepresented and disadvantaged demographic groups at inference time (Barocas & Selbst, 2016; Bolukbasi et al., 2016). As a result, concerns of fairness have gained significant attention, especially as applications of these models expand to high-stakes domains such as criminal justice, healthcare, and finance (Berk et al., 2021). To mitigate the bias, a variety of fairness criteria and algorithms have been proposed (Barocas et al., 2023; Caton & Haas, 2020), which impose mathematical or statistical constraints on the model to ensure equitable treatment under the respective fairness notions. But these algorithms typically incur a cost to model performance as they improve model fairness (Calders et al., 2009; Corbett-Davies et al., 2017).

It is not immediately clear whether the degradation in 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 & Gordon, 2022). Hence the design of fair algorithms necessitates the understanding of this tradeoff, which would also provide insight to the implications of fairness in machine learning; yet, it remains an open problem for most fairness criteria and learning 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) concurrently characterized the tradeoff between mean squared error (MSE) and 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. (2023) assumed binary groups, Zeng et al. (2022a) and Gaucher et al. (2023) assumed binary class labels, and Zhao & Gordon (2022) assumed that the data distribution is *noiseless*, i.e., the Bayes error rate is zero. We will close this gap and complete 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 (approximate) DP fairness in the general setting of multi-group, multi-class, and potentially *noisy* data distributions. 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 characterization reveals that the optimal fair classifier—one that satisfies DP while achieving the minimum error—is given by the composition of the Bayes optimal score function (minimum MSE regressor of the one-hot labels) and the optimal transports from the Wasserstein-barycenter problem (Section 3.2).
3. Based on the findings, we propose a post-processing

[1]University of Illinois Urbana-Champaign, Urbana, Illinois, USA. Correspondence to: Ruicheng Xian <rxian2@illinois.edu>, Han Zhao <hanzhao@illinois.edu>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

method that derives fair classifiers from (pre-trained) score functions (Section 3.3). Our method is instantiated for finite sample estimation in Section 4 with sample complexity analysis.[2]

4. Experiments on benchmark datasets demonstrate the effectiveness of our algorithm (Section 5), which achieves precise control of the tradeoff provided sufficient training data.

1.1. Related Work

Inherent Tradeoff. The concept of barycenter appears in many analyses of the tradeoff of DP fairness. Intuitively, by treating the barycenter—computed over the distributions of optimal model outputs (without constraints) on each group—as the output distribution that is required to be identical across groups under DP, the sum of distances to the barycenter is naturally related to the minimum fair error.

We review existing characterizations of the tradeoff of DP fairness below and draw connections to our result. Denote the input by X, group membership by A, and target variable by Y (for classification, the *one-hot* label). Let r_a^* be the *distribution* of the conditional mean on group a, $\mathbb{E}[Y \mid X, A = a]$, i.e., the minimum MSE estimates of Y given $(X, A = a)$ (for classification, these are *distributions* of class probabilities). Lastly, let $w_a \coloneqq \mathbb{P}(A = a)$ denote the proportion of each group. Then, under DP, on

- regression problems (Le Gouic et al., 2020; Chzhen et al., 2020; Chzhen & Schreuder, 2022), the minimum *excess* risk in terms of MSE is given by the Wasserstein-2-barycenter (under the ℓ_2 metric) over the r_a^*'s,

$$\min_{q:\mathrm{supp}(q)\subseteq\mathbb{R}} \sum_{a\in\mathcal{A}} w_a \, W_2^2(r_a^*, q); \tag{1}$$

- *noiseless* classification problems (Zhao & Gordon, 2022), the minimum/excess error rate is given by the TV-barycenter over the class priors, $p_a(e_i) \coloneqq \mathbb{P}(Y = e_i \mid A = a)$,

$$\min_{q:\mathrm{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \sum_{a\in\mathcal{A}} \frac{w_a}{2} \|p_a - q\|_1, \tag{2}$$

  where $\frac{1}{2}\|\cdot\|_1$ computes the total variation (TV);

- classification problems in the general setting (Theorem 3.2), the minimum error rate is given by the Wasserstein-1-barycenter (under the ℓ_1 metric),

$$\min_{q:\mathrm{supp}(q)\subseteq\{e_1,\cdots,e_k\}} \sum_{a\in\mathcal{A}} \frac{w_a}{2} \, W_1(r_a^*, q). \tag{3}$$

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

First, unlike regression, the support of the barycenters in Equations (2) and (3) is restricted to $\{e_1, \cdots, e_k\}$, which represents 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 class *assignments*, the minimum error rate equals the sum of $\frac{1}{2}W_1$ distances to the barycenter under the ℓ_1 metric. Similarly, the use of the W_2^2 distance under ℓ_2 in Equation (1) reflects the MSE loss. Second, our Equation (3) recovers Equation (2) in the noiseless setting, because, under which, $r_a^* = p_a$ and $\frac{1}{2}W_1 = \frac{1}{2}\|\cdot\|_1$. Denis et al. (2023) and Gaucher et al. (2023) also derived similar expressions for the tradeoff to ours, but only under binary group or class labels.

Post-Processing. Given a (biased) model, this family of mitigation algorithms post-process the model to satisfy fairness, e.g., via remapping the outputs (Hardt et al., 2016; Pleiss et al., 2017). Existing algorithms for DP fairness include (Fish et al., 2016; Menon & Williamson, 2018; Chzhen et al., 2019; Jiang et al., 2020; Zeng et al., 2022a; Denis et al., 2023), but they are limited to binary group and/or binary classification.

For multi-group and multi-class DP, the only applicable post-processing algorithm, to our knowledge, is due to Alghamdi et al. (2022), which is based on model projection. But the tradeoff of their algorithm is unclear as they did not directly relate error rate to the difference between the projected model and the original, and experiments show that their algorithm underperforms compared to ours, especially on tasks involving a large number of groups and classes.

2. Preliminaries

Notation. Denote the $(k-1)$-dimensional probability simplex by $\Delta_k \coloneqq \{z \in \mathbb{R}^k_{\geq 0} : \|z\|_1 = 1\}$, whose k vertices are $\{e_1, \cdots, e_k\}$, where $e_i \in \mathbb{R}^k$ is the vector of all zeros except for 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 work with *randomized functions* (Definition B.2), which have probabilistic outputs according to some distributions conditioned on the input. Given a (randomized) function $f : \mathcal{X} \to \mathcal{Y}$ and a distribution p over $\mathcal{X}$, we denote the *push-forward* of p by $f\sharp p$ (Definition B.3).

Problem Setup. A k-class classification problem is defined by a joint distribution μ of input $X \in \mathcal{X}$, demographic group membership (a.k.a. the sensitive attribute) $A \in \mathcal{A} = [m] \coloneqq \{1, \cdots, m\}$, and class label in one-hot representation, $Y \in \mathcal{Y} = \{e_1, \cdots, e_k\}$; the class labels may be subject to noise originating from, e.g., the data collection process. Denote the marginal distribution of input X by μ^X, the conditional distribution of μ on group $A = a$ by μ_a, and the group weight by $w_a \coloneqq \mathbb{P}_\mu(A = a)$.

The goal of fair classification is to find a *randomized* and *attribute-aware* classifier, $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$, that achieves the minimum classification error rate on μ subject to the constraints set by the designated fairness criteria. Denote the component of h associated with group a by $h_a : \mathcal{X} \to \mathcal{Y}$, i.e., $h_a(x) \equiv h(x, a)$. The error rate is defined as

$$\begin{aligned}\mathrm{err}(h) &\coloneqq \mathbb{P}(h_A(X) \neq Y) \\ &= \sum_{a\in[m]} w_a \, \mathbb{P}(h_A(X) \neq Y \mid A = a) \\ &= \sum_{a\in[m]} w_a \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{P}(h_a(x) \neq y) \, \mathrm{d}\mu_a(x, y),\end{aligned}$$

where the decomposition on the last line highlights the randomness of h. For fairness, we consider the group criterion of demographic parity:

Definition 2.1 (Approximate Demographic Parity)**.** For $\alpha \in [0, 1]$, a classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ is said to satisfy α-DP if $\Delta_{\mathrm{DP}}(h) \leq \alpha$, which is defined as

$$\begin{aligned}\Delta_{\mathrm{DP}}(h) &\coloneqq \max_{\substack{a,a'\in[m] \\ y\in\mathcal{Y}}} \big|\mathbb{P}(h_A(X) = y \mid A = a) \\ &\qquad\qquad - \mathbb{P}(h_A(X) = y \mid A = a')\big| \\ &= \max_{a,a'\in[m]} \big\|h_a\sharp\mu_a^X - h_{a'}\sharp\mu_{a'}^X\big\|_\infty,\end{aligned}$$

where

$$\mathbb{P}(h_A(X) = y \mid A = a) = \int_{\mathcal{X}} \mathbb{P}(h_a(x) = y) \, \mathrm{d}\mu_a^X(x)$$

is the proportion of outputs with class assignment y on group a, and $\|p - q\|_\infty \coloneqq \max_{z\in\mathcal{Z}} |p(z) - q(z)|$ between two distributions p, q.

We call a classifier α-*fair* if it satisfies α-DP. The parameter α controls the tradeoff between fairness and (the maximum attainable) accuracy (due to the inherent tradeoff); setting $\alpha = 0$ recovers the standard strict definition of DP.

Lastly, a (attribute-aware) score function $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ is a model that outputs probability vectors as estimates of the class probabilities, as in $f(x, a)_y \approx \mathbb{P}_{\mu_a}(Y = y \mid X = x)$. A score function is said to be *Bayes optimal*, denoted by f^*, if it computes the true class probabilities exactly,

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

it coincides with the minimum MSE estimator of the one-hot labels Y given $(X, A = a)$. We will often work with the quantity $r_a^* \coloneqq f_a^*\sharp\mu_a^X$, the distribution of true class probabilities conditioned on group a.

Given a (pre-trained) score function f, our post-processing method finds a (probabilistic) fair classifier by deriving from f. I.e., it returns classifiers of the form $(x, a) \mapsto g_a \circ f_a(x)$ for some post-processing maps $g_1, \cdots, g_m : \Delta_k \to \mathcal{Y}$.

Optimal Transport and Wasserstein Distance. Our analysis involves the concept of optimal transports and Wasserstein distance (Villani, 2003); the latter is a metric on the space of probability distributions.

Definition 2.2 (Coupling)**.** Let p, q be probability distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively. A coupling γ of p, q is a joint distribution over $\mathcal{X} \times \mathcal{Y}$ satisfying $p(x) = \int_{y\in\mathcal{Y}} \mathrm{d}\gamma(x, y)$, $\forall x \in \mathcal{X}$, and $q(y) = \int_{x\in\mathcal{X}} \mathrm{d}\gamma(x, y)$, $\forall y \in \mathcal{Y}$. We denote the collection of couplings of p, q by $\Gamma(p, q)$.

Definition 2.3 (Optimal Transport)**.** Let p, q be probability distributions over $\mathcal{X}$ and $\mathcal{Y}$, respectively, and $c : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ a cost function. The optimal transportation cost between p and q is given by $\inf_{\gamma\in\Gamma(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} c(x, y) \, \mathrm{d}\gamma(x, y)$. Let γ^* be a minimizer, then the optimal transport from p to q, denoted by $\mathcal{T}^*_{p\to q,c} : \mathcal{X} \to \mathcal{Y}$, is a (randomized) function satisfying $\gamma^* = (\mathrm{Id} \times \mathcal{T}^*_{p\to q,c})\sharp p$, where Id is the identity map (that in the other direction is defined symmetrically).

Intuitively, $\mathcal{T}^*_{p\to q,c}$ specifies a plan for moving masses distributed according to p to q with the minimum total cost. In this plan, the mass located at each $x \in \mathcal{X}$ is moved (probabilistically) to $\mathcal{T}^*_{p\to q,c}(x) \in \mathcal{Y}$. The optimal transport can also be represented by the optimal coupling $\gamma^* \in \Gamma(p, q)$, as we can derive an optimal transport $\mathcal{T}$ from γ^* by setting $\mathbb{P}(\mathcal{T}(X) = y \mid X = x) = \gamma^*(x, y)/\gamma^*(x, \mathcal{Y})$, $\forall x, y$.[3]

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

Definition 2.4 (Wasserstein Distance)**.** Let p, q be probability distributions over 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

In this section, we provide a characterization of the inherent tradeoff of DP fairness, then, based on the findings, propose and analyze a post-processing method for DP.

3.1. Characterization of the Inherent Tradeoff

Our characterization comes from a reformulation of the classification problem assuming access to the Bayes optimal score. On any generic (group-less) classification problem,

Lemma 3.1. *Let $f^* : \mathcal{X} \to \Delta_k$ be the Bayes optimal score function, define $r^* \coloneqq f^*\sharp\mu^X$, and fix $q \in \mathcal{Q}_k$. 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. $\mathrm{err}(h) = \frac{1}{2}\int_{\Delta_k\times\mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma(s, y)$. Conversely, for any $\gamma \in \Gamma(r^*, q)$, there exists a randomized classifier h satisfying $h\sharp\mu^X = q$ s.t. the above equality holds.*

[3] 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.

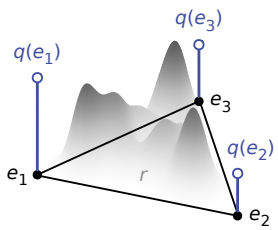

Figure 1. A distribution r over the 2d simplex (density; grey surface), and a finite distribution $q \in \mathcal{Q}_3$ over its vertices $\{e_1, e_2, e_3\}$ (blue spikes).

It follows that minimizing classification error subject to having an output distribution of q is equivalent to solving an optimal transport problem from r^* to q under the ℓ_1 cost, because by Definition 2.4, $\min_{h:h\sharp\mu^X=q} \mathrm{err}(h) = \frac{1}{2}\min_{\gamma\in\Gamma(r^*,q)} \int \|s-y\|_1 \,\mathrm{d}\gamma = \frac{1}{2} W_1(r^*, q)$.

Note that this reformulation allows for explicit control of the output distribution, which is well-suited for analyzing DP fairness since it only constrains the output distributions $q_a \coloneqq h_a\sharp\mu_a^X$; namely, that they need to be equal ($\alpha = 0$) or close: $\Delta_{\mathrm{DP}}(h) \le \alpha \iff \max_{a,a'} \|q_a - q_{a'}\|_\infty \le \alpha$ by Definition 2.1. Therefore, for attribute-aware classifiers, whose component h_a's can be optimized independently, the discussions above immediately give the following characterization of the minimum error rate under DP:

Theorem 3.2 (Minimum Fair Error Rate)**.** *Let $\alpha \in [0,1]$, $f^* : \mathcal{X}\times\mathcal{A} \to \Delta_k$ be the Bayes optimal score function, and define $r_a^* \coloneqq f_a^*\sharp\mu_a^X$, $\forall a \in [m]$. With W_1 under the ℓ_1 metric,*

$$\begin{aligned}\mathrm{err}_\alpha^* &\coloneqq \min_{h:\Delta_{\mathrm{DP}}(h)\le\alpha} \mathrm{err}(h)\\ &= \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k\\ \max_{a,a'}\|q_a-q_{a'}\|_\infty\le\alpha}} \sum_{a\in[m]} \frac{w_a}{2} W_1(r_a^*, q_a). \quad (4)\end{aligned}$$

It could be viewed as a (relaxed) *Wasserstein-barycenter* problem on the r_a^*'s under the special case where the support of the barycenter(s) q_a is restricted to the vertices $\{e_1, \cdots, e_k\}$. It is a convex problem (in the primal form presented above), and can be simplified under certain assumptions: if the problem is noiseless (and $\alpha = 0$), it reduces to the TV-barycenter problem in Equation (2) (Theorem A.1); this result is first established in (Zhao & Gordon, 2022), but only under $m = k = 2$.

Under strict DP fairness ($\alpha = 0$), the inherent tradeoff, namely the excess risk incurred by the DP constraint, is $\frac{1}{2}(\min_q \sum_a w_a W_1(r_a^*, q) - \sum_a \min_{q_a} w_a W_1(r_a^*, q_a)) \ge 0$; the second term is the Bayes error rate, achieved by the classifier $(x, a) \mapsto e_{\arg\max_i f_a^*(x)_i}$. The tradeoff is expected to be large on problems with very different r_a^*'s, and equals to zero when they are identical (i.e., $\mathbb{E}_\mu[Y \mid X, A] \perp\!\!\!\perp A$; since all groups would have the same optimal decision rule), meaning that enforcing DP would not degrade model performance. But we point out that the tradeoff could be zero even if $\mathbb{E}_\mu[Y \mid X, A] \not\perp\!\!\!\perp A$, partly due to the nonuniqueness of the optimal classifier (Example A.4).

Lastly, Zhao & Gordon (2022) concluded that in the noiseless setting, the tradeoff is zero if and only if the class priors are identical, i.e., $\mathbb{E}_\mu[Y \mid A] \perp\!\!\!\perp A$. But this condition is no longer sufficient in the general setting (Example A.5).

3.2. Optimal Fair Classifier via Post-Processing

In addition to characterizing the minimum error rate under DP, we also show that the optimal fair classifier can be obtained by deriving from the Bayes optimal score f^*:

Theorem 3.3 (Optimal Fair Classifier)**.** *Let $\alpha \in [0,1]$, $f^* : \mathcal{X}\times\mathcal{A} \to \Delta_k$ be the Bayes optimal score function, $(q_1^*, \cdots, q_m^*)$ a minimizer of Equation (4), and $\mathcal{T}^*_{r_a^*\to q_a^*}$ the optimal transport from r_a^* to q_a^* under the ℓ_1 cost, $\forall a \in [m]$. We have*

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

This result is a consequence of the construction used in the proof of Lemma 3.1 (deferred to Appendix B), and reveals the form of the optimal fair classifier as a composition of f^* and optimal transports from r_a^*'s to the minimizing q_a^*'s of Equation (4) (see Figure 1 for a picture of these distributions). It immediately suggests a three-step method for learning optimal fair classifiers: *(i)* learn the Bayes optimal score function f^*, e.g., via minimizing MSE w.r.t. the one-hot label Y,

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

(ii) find a minimizer $(q_1^*, \cdots, q_m^*)$ of the barycenter problem in Equation (4), *(iii)* compute the optimal transports $\mathcal{T}^*_{r_a^*\to q_a^*}$, and finally return $(x, a) \mapsto \mathcal{T}^*_{r_a^*\to q_a^*} \circ f_a^*(x)$. The last two steps (reproduced in Algorithm 1) post-process f^*.

3.3. Post-Processing Any Score Function

In practice, however, the Bayes optimal f^* may not be exactly learned due to computational cost, or difficulties in representation, optimization, and generalization (Woodworth et al., 2017). Instead, we will often work with suboptimal scores $f \approx f^*$ (e.g., pre-trained by a vendor). This section analyzes the applicability and suboptimality of Algorithm 1 for post-processing non-Bayes optimal score functions.

Given an arbitrary score function $f : \mathcal{X}\times\mathcal{A} \to \Delta_k$, we want to find post-processing maps $g_a : \Delta_k \to \mathcal{Y}$ such that the derived classifier $(x, a) \mapsto g_a \circ f_a(x)$ satisfies DP fairness,

Algorithm 1 Post-Process for α-DP

Input: $\alpha \in [0,1]$, score function $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$, marginal distribution $\mu^{X,A}$ of (X, A)
Define $w_a \coloneqq \mathbb{P}_\mu(A = a)$ and $r_a \coloneqq f_a \sharp \mu_a^X$, $\forall a \in [m]$
$(q_1, \cdots, q_m) \leftarrow$ minimizer of Eq. (4)
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: $(x, a) \mapsto \mathcal{T}^*_{r_a \to q_a} \circ f_a(x)$

and ideally, achieves the minimum error rate among all fair classifiers derived from f.

Let $\bar{h}(x, a) \coloneqq \mathcal{T}^*_{r_a \to q_a} \circ f_a(x)$ denote the classifier obtained from applying Algorithm 1 to f. First, $\bar{h}$ is always α-fair regardless of f, because $\bar{h}_a \sharp \mu_a^X = q_a$, $\forall a \in [m]$ by construction, and $\Delta_{\mathrm{DP}}(\bar{h}) = \max_{a,a'} \|q_a - q_{a'}\|_\infty \le \alpha$ from the constraints in Equation (4). Second, the suboptimality of $\bar{h}$ can be upper bounded by the L^1 difference between f and the Bayes optimal score f^*:

Theorem 3.4 (Error Propagation)**.** *Let $\alpha \in [0,1]$, $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ be a score function, and f^* the Bayes optimal score function. For the α-fair classifier $\bar{h}$ obtained from applying Algorithm 1 to f,*

$$0 \le \mathrm{err}(\bar{h}) - \mathrm{err}^*_\alpha \le \mathbb{E}[\|f(X, A) - f^*(X, A)\|_1],$$

where err^*_α is defined in Equation (4).

Hence, whereas the degradation in performance of the post-processed $\bar{h}$ from Algorithm 1 if $f = f^*$ is attributed entirely to the inherent tradeoff (Theorem 3.3), it is not the case when $f \neq f^*$ due to the loss of information about Y. We may, however, guarantee that $\bar{h}$ is optimal among all fair classifiers derived from f if it satisfies *group-wise distribution calibration* (Kull & Flach, 2015), i.e., the output predictions correctly convey the class probabilities:

Definition 3.5. A score function $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ is said to be group-wise distribution calibrated if $\mathbb{P}_\mu(Y = e_i \mid f(X, a) = s, A = a) = s_i, \forall s \in \Delta_k, i \in [k], a \in [m]$.

If f is not calibrated, but labeled data is available, one could learn mappings $u_a : \Delta_k \to \Delta_k$ and compose it with f to recalibrate it. The optimal calibration maps, u_a^* (in the sense that they achieve calibration without incurring further information loss), are by definition the minimum MSE estimators of Y given $(f_a(X), A = a)$.

To see why the optimality of $\bar{h}$ among all fair classifiers derived from f is guaranteed provided calibration, note that finding the optimal fair *post-processing map* for f is equivalent to finding the optimal fair *classifier* on a new problem μ' derived from the original μ under an input transformation, given by the joint distribution of $(X' \coloneqq f_A(X), A, Y)$. Also, the Bayes optimal score on μ' coincide with the optimal calibration map, as $\mathbb{E}_{\mu'}[Y \mid X' = s, A = a] = u_a^*(s)$. So by Theorem 3.3, Algorithm 1 finds post-processing map g_a's s.t. $(x', a) \mapsto g_a \circ u_a^*(x')$ is the optimal fair classifier on μ', whereby $(x, a) \mapsto g_a \circ (u_a^* \circ f_a)(x)$ is optimal among all derived fair classifiers (the term in the parentheses is the recalibrated score). Finally, we remark that the suboptimality of $\bar{h}$ due to miscalibration can be bounded using Theorem 3.4 by simply replacing the reference f^* with the calibrated score $(x, a) \mapsto u_a^* \circ f_a$.

4. Finite Sample Estimation

We have discussed post-processing for DP assuming access to the distribution $\mu^{X,A}$. In this section, we instantiate our Algorithm 1 for post-processing using finite samples:

Assumption 4.1. We have n i.i.d. samples of (X, A) that are independent of the score function f being post-processed.

Denote the samples from group a by $(x_{a,i})_{i \in [n_a]}$, and their number by n_a. Define $\hat{w}_a \coloneqq n_a/n$, and the empirical distribution $\hat{r}_a \coloneqq \frac{1}{n_a} \sum_{i \in [n_a]} \delta_{f_a(x_{a,i})}$, where δ is the Dirac delta function.

We also analyze the sample complexity for both the fairness and the error rate of the returned classifier. For simplicity, we assume f to be calibrated; otherwise, Theorem 3.4 can be used to bound the suboptimality due to miscalibration.

Assumption 4.2. The score function f being post-processed is *group-wise calibrated* (Definition 3.5).

Let $r_a \coloneqq f_a \sharp \mu_a^X$, $\forall a \in [m]$. Recall from Section 3.3 that the error rate of the optimal derived α-fair classifier is

$$\mathrm{err}^*_{\alpha,f} \coloneqq \min_{\substack{q_1, \cdots, q_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q_a - q_{a'}\|_\infty \le \alpha}} \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a, q_a).$$

This section is divided into three subsections w.r.t. the continuity of the distributions of the score, $r_1, \cdots, r_m$.

4.1. The Finite Case

We start with the case where the r_a's have finite supports, i.e., $|\mathcal{R}_a| < \infty$ where $\mathcal{R}_a \coloneqq \mathrm{supp}(r_a)$. Note that this does not mean that the input distribution μ^X is finite.

If the true probability mass of the r_a's were known, then Algorithm 1 can be implemented by a linear program:

$$\begin{aligned}
\mathrm{LP}: \quad & \min_{\substack{q_1, \cdots, q_m \ge 0 \\ \gamma_1, \cdots, \gamma_m \ge 0}} \sum_{a \in [m]} \sum_{s \in \mathcal{R}_a, y \in \mathcal{Y}} \frac{w_a}{2} \|s - y\|_1 \gamma_a(s, y) \\
& \text{s.t.} \quad \sum_{s' \in \mathcal{R}_a} \gamma_a(s', y) = q_a(y), \quad \forall a \in [m],\ y \in \mathcal{Y}, \\
& \qquad \sum_{y' \in \mathcal{Y}} \gamma_a(s, y') = r_a(s), \quad \forall a \in [m],\ s \in \mathcal{R}_a, \\
& \qquad |q_a(y) - q_{a'}(y)| \le \alpha, \quad \forall a, a' \in [m],\ y \in \mathcal{Y},
\end{aligned}$$

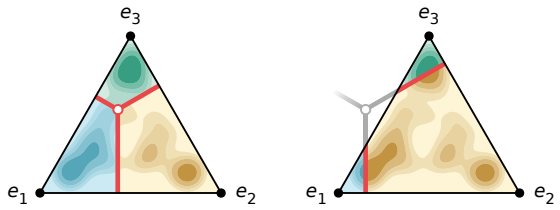

Figure 2. Examples of simplex-vertex optimal transports ($k = 3$; with different vertex distributions). All points in the lower-left blue partition are transported to e_1, lower-right yellow to e_2, and upper green to e_3. The transports are described by a Y-shaped boundary.

where $q_a \in \Delta_k$ and $\gamma_a \in \mathbb{R}^{|\mathcal{R}_a| \times k}$. This program simultaneously finds a minimizer $(q_1^*, \cdots, q_m^*)$ of the barycenter problem in Equation (4) and the optimal transports $\mathcal{T}^*_{r_a \to q_a^*}$ (used in Theorem 3.3) in the form of couplings $(\gamma_1^*, \cdots, \gamma_m^*)$: namely, 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 true pmfs of the r_a's are unknown but finite samples as in Assumption 4.1 are given, we proceed with solving LP defined on the empirical $\hat{w}_a$ and $\hat{r}_a$'s, which will give us estimated $\hat{q}_a$'s and $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$'s. Then, we post-process f via $\hat{h}(x, a) = \mathcal{T}^*_{\hat{r}_a \to \hat{q}_a} \circ f_a(x)$. The sample complexity is:

Theorem 4.3 (Sample Complexity, Finite Case)**.** *Let $\alpha \in [0, 1]$, $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ be a score function, and assume $|\mathcal{R}_a| \coloneqq \mathrm{supp}(f_a \sharp \mu_a^X) < \infty$, $\forall a \in [m]$. W.p. at least $1 - \delta$ over the random draw of samples in Assumption 4.1, for the classifier $\hat{h}$ derived above, and $n \geq \Omega(\max_a \ln(m/\delta)/w_a)$,*

$$\Delta_{\mathrm{DP}}(\hat{h}) \leq \alpha + O\left(\max_a \sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}}\right);$$

in addition, with Assumption 4.2,

$$\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\alpha, f} \leq O\left(\max_a \sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}}\right).$$

4.2. The Continuous Case

When the r_a's are continuous,[4] given finite samples, we may still solve LP defined on $\hat{w}_a$ and $\hat{r}_a$'s to estimate the optimal output distribution $\hat{q}_a$'s under α-DP, but the empirical transports $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ are no longer usable for post-processing in this case, since by continuity, the inputs to the transports at inference time will be unseen almost surely (i.e., $f_a(x) \notin f_a[(x_{a,i})_{i \in [n_a]}]$ a.s. for $x \sim \mu_a^X$), or in other words, they cannot extrapolate to the full support of r_a. So after obtaining the $\hat{q}_a$'s, we will need to estimate the optimal transports $\mathcal{T}^*_{r_a \to \hat{q}_a}$ from (the population) r_a's to the $\hat{q}_a$'s.

Since $\mathrm{supp}(\hat{q}_a) = \{e_1, \cdots, e_k\}$ is finite, this makes finding $\mathcal{T}^*_{r_a \to \hat{q}_a}$ a semi-discrete optimal transport problem (Genevay et al., 2016; Staib et al., 2017; Chen et al., 2019), for which, a common procedure in existing work is to reformulate 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: concretely, for each $a \in [m]$,

$$\begin{aligned} W_1(r_a, \hat{q}_a) &= \inf_{\gamma_a \in \Gamma(r_a, \hat{q}_a)} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma_a(s, y) \\ &= \sup_{\substack{\phi_a : \Delta_k \to \mathbb{R},\, \psi_a \in \mathbb{R}^k \\ \phi_a(s) + \psi_{a,i} \leq \|s - e_i\|_1}} \left(\mathbb{E}_{S \sim r_a}[\phi_a(S)] + \sum_{i=1}^k \psi_{a,i} \hat{q}_a(e_i) \right) \\ &= \sup_{\psi_a \in \mathbb{R}^k} \left(\mathbb{E}_{S \sim r_a}\left[\min_i (\|S - e_i\|_1 - \psi_{a,i})\right] + \sum_{i=1}^k \psi_{a,i} \hat{q}_a(e_i) \right), \end{aligned} \tag{5}$$

and ψ_a^* can be optimized, e.g., using (stochastic) gradient ascent. Moreover, Gangbo & McCann (1996) showed that in the semi-discrete case, the optimal transport $\mathcal{T}^*_{r_a \to \hat{q}_a}$ belongs to the parameterized function class

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

(break ties to the tied e_i with the largest index i) with ψ_a^* as its parameter. See Figure 2 for pictures of semi-discrete optimal transports when $k = 3$. Compared to the empirical transports $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$, the domain of $\mathcal{T}^*_{r_a \to \hat{q}_a} \in \mathcal{G}_k$ covers the support of r_a, i.e., it can handle future unseen inputs.

Note that Gangbo & McCann (1996) assumed 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 & Pratelli, 2003). Hence, as a technical contribution, we provide a proof in Appendix D for the existence and uniqueness of the optimal transport in $\mathcal{G}_k$ on simplex-vertex transportation problems under the ℓ_1 cost, via analyzing its geometry.

Our Implementation. To recap, for post-processing in the finite sample and continuous case, we *(i)* get the $\hat{q}_a$'s from solving LP, then *(ii)* estimate the $\mathcal{T}^*_{r_a \to \hat{q}_a}$'s; as discussed above, the typical approach for solving this semi-discrete transportation problem is via optimizing ψ_a w.r.t. Equation (5) (taking expectation over the empirical $\hat{r}_a$).

However, instead of estimating from scratch, by leveraging the observation that each of the empirical transports $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ obtained (as byproducts) from solving LP already achieves the optimal value of $W_1(\hat{r}_a, \hat{q}_a)$ of Equation (5), and their

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

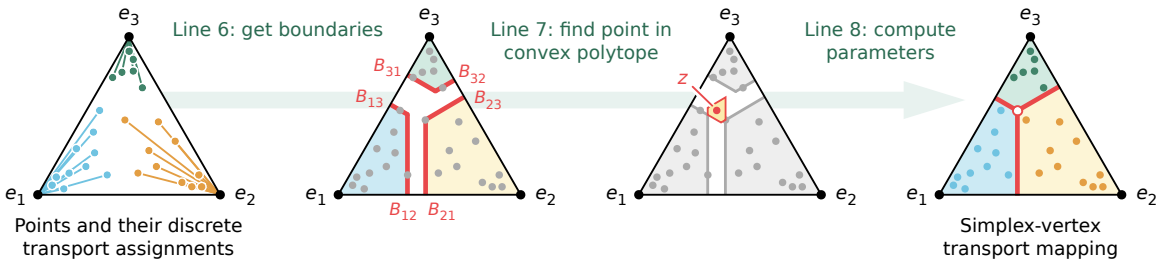

Figure 3. Extract a simplex-vertex transport $\mathcal{T} \in \mathcal{G}_3$ that agrees with the discrete optimal transport (Lines 6 to 9 of Algorithm 2). For illustrative purposes, the discrete transport on the left does not split mass; otherwise, there would be disagreements on the boundaries.

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

1: **Input:** $\alpha \in [0,1]$, score function $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$, samples $((x_{a,i})_{i\in[n_a]})_{a\in[m]}$
2: Define $\hat{w}_a := \frac{n_a}{n}$ and $\hat{r}_a := \frac{1}{n_a}\sum_i \delta_{f_a(x_{a,i})}, \forall a \in [m]$
3: $(\gamma_1, \cdots, \gamma_m) \leftarrow$ minimizer of LP on $\hat{w}_a$ and $\hat{r}_a$'s
4: Define $v_{ij} := e_j - e_i$
5: **for** $a = 1$ **to** m **do**
6: $\quad B_{a,ij} \leftarrow \{0\} \cup \max\{f_a(x_{a,\ell})^\top v_{ij} + 1 : \ell \text{ s.t. } \gamma_a(f_a(x_{a,\ell}), e_i) > 0\}$
7: $\quad z_a \leftarrow$ point in $\bigcap_{i\neq j}\{x \in \mathbb{R}^k : x^\top v_{ij} \geq B_{a,ij} - 1\}$
8: $\quad \psi_{a,i} \leftarrow 2z_a^\top v_{i1}, \forall i \in [k]$
9: $\quad \mathcal{T}_a \leftarrow (s \mapsto e_{\arg\min_i(\|s-e_i\|_1 - \psi_{a,i})})$
10: **end for**
11: **Return:** $(x, a) \mapsto \mathcal{T}_a \circ f_a(x)$

decision boundaries is simply described by $\binom{k}{2} = \Theta(k^2)$ hyperplanes (recall Figure 2), we can directly extract a set of transport mappings $\mathcal{T}_a \in \mathcal{G}_k$ from $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$. The procedure for this is provided on Lines 6 to 9 in Algorithm 2 (with step-by-step illustration in Figure 3; formal derivations are deferred to Appendix D), which amounts to finding a feasible point in a polytope, formulated as a linear program (Appendix E.3). Each of the extracted $\mathcal{T}_a$ will agree with $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$ on all points in $\hat{r}_a$ except for those that lie on the $\Theta(k^2)$ boundaries, and by continuity of r_a, the number of disagreements between them is $O(k^2)$ almost surely.

Our implementation in Algorithm 2 involves $(m+1)$ linear programs, where LP dominates with $O(nk)$ variables and constraints, and takes, e.g., $\widetilde{O}(\text{poly}(nk))$ time to solve to (near-)optimality using interior point methods (Vaidya, 1989). Its sample complexity is:

Theorem 4.4 (Sample Complexity, Continuous Case)**.** *Let $\alpha \in [0,1]$, $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ be a score function, and assume that $f_a \sharp \mu_a^X$ is continuous, $\forall a \in [m]$. W.p. at least $1-\delta$ over the random draw of samples in Assumption 4.1, for the classifier $\hat{h}$ obtained from applying Algorithm 2 to f, and $n \geq \Omega(\max_a \ln(m/\delta)/w_a)$,*

$$\Delta_{\text{DP}}(\hat{h}) \leq \alpha + O\left(\max_a\left(\sqrt{\frac{k + \ln(mk/\delta)}{nw_a}} + \frac{k}{nw_a}\right)\right);$$

in addition, with Assumption 4.2,

$$\text{err}(\hat{h}) - \text{err}^*_{\alpha,f} \leq O\left(\max_a\left(\sqrt{\frac{k\ln(m/\delta)}{nw_a}} + \frac{k^2}{nw_a}\right)\right).$$

The first term in both expressions is the sample complexity of PAC learning (with the complexity of $\mathcal{G}_k$ analyzed in Theorem D.2), and the second term comes from the disagreement between $\mathcal{T}_a$ and $\mathcal{T}^*_{\hat{r}_a \to \hat{q}}$ on $\hat{r}_a$ discussed above.

4.3. The General Case

For completeness, we briefly discuss the general case where the r_a's are neither finite nor continuous (i.e., contain atoms), which is handled by smoothing the r_a's using an i.i.d. noise generator ρ with a continuous distribution.[4]

The smoothing is done by perturbing (samples from) r_a with random noise drawn from ρ, i.e., $\tilde{r}_a := u_\rho \sharp r_a$, where u_ρ is a randomized function s.t. $u_\rho(s) \sim s + N$ with $N \sim \rho$; it is not hard to show that the resulting $\tilde{r}_a$ is continuous. Now that $u_\rho \circ f_a$ produces continuous score distributions, we may apply the same algorithm in Section 4.2 for DP post-processing. The resulting classifier, $\bar{h}_\rho(x,a) := \mathcal{T}^*_{\tilde{r}_a \to q_a} \circ u_\rho \circ f_a(x)$, is α-fair regardless of the choice of ρ, and the suboptimality incurred by the smoothing procedure is controlled by the *bandwidth* of ρ:

Theorem 4.5 (Error Propagation, Smoothing)**.** *Let $\alpha \in [0,1]$, $f : \mathcal{X} \times \mathcal{A} \to \Delta_k$ be a score function, and ρ a continuous distribution with finite first moment. Under Assumption 4.2, for the α-fair classifier $\bar{h}_\rho$ derived above,*

$$0 \leq \text{err}(\bar{h}_\rho) - \text{err}^*_{\alpha,f} \leq \mathbb{E}_{N\sim\rho}[\|N\|_1].$$

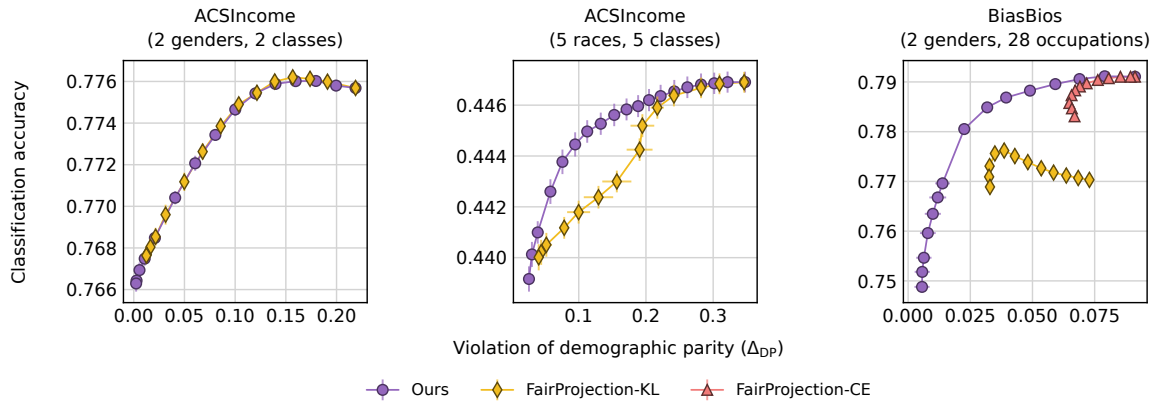

Figure 4. Tradeoff curves between accuracy and Δ_{DP} (Definition 2.1). Scoring model is logistic regression. Error bars indicate the standard deviation over 10 runs with different random splits. Running time is reported in appendix Table 1. On ACSIncome, `FairProjection-KL` and `FairProjection-CE` have similar results.

E.g., if $\rho = \text{Laplace}(0, b\, I_k)$, then $\mathbb{E}[\|N\|_1] = kb$, and the suboptimality due to smoothing is less than kb; in practice, the smallest-allowable b depends on machine precision.

5. Experiments

Our proposed DP post-processing Algorithm 2 is evaluated on four benchmark datasets: the UCI Adult dataset for income prediction (Kohavi, 1996), the ACSIncome dataset (Ding et al., 2021)—an extension of the Adult with much more examples (1.6 million vs. 48,842)—on which we consider a binary setting where the sensitive attribute is gender and the target is whether the income is over \$50k, as well as a multi-group multi-class setting with five race categories and five income buckets; the Communities & Crime dataset (Redmond & Baveja, 2002), and the BiasBios dataset (De-Arteaga et al., 2019), where the task is to predict occupations from biographies. We highlight the effectiveness of our algorithm by comparing it to `FairProjection` (Alghamdi et al., 2022), which is to our knowledge the only other fair post-processing algorithm for multi-group and multi-class classification.

On each dataset, we split the data into pre-training, post-processing, and testing. We first train a linear logistic regression scoring model on the pre-training split, then perform DP post-processing. On BiasBios, the model is trained on embeddings of biographies computed by a pre-trained BERT language model from the `bert-base-uncased` checkpoint (Devlin et al., 2019). Additional details, including hyperparameters, are in Appendix E.

Results. The results on ACSIncome and BiasBios are shown in Figure 4 (those on Adult and Communities are deferred to appendix Figure 7). Across all tasks, our method is effective at reducing the disparity, and almost precise control of Δ_{DP} via α is achieved on tasks with sufficient data. Compared to `FairProjection`, our method can achieve lower Δ_{DP} and produces better tradeoff curves, and its advantage is most evident under multi-class settings (e.g., BiasBios).

While our method achieves a significant degree of DP fairness, there remains a gap to $\Delta_{\text{DP}} = 0$ in our results, especially on tasks with more groups and classes (e.g., ACSIncome under the multi-group multi-class setting). This could be due to a potential violation of the continuity assumption required by Algorithm 2 (Section 4.2), but we suspect the main reason to be insufficient data. Recall from Theorem 4.4 that the sample complexity scales as $\widetilde{O}(\sqrt{k/nw_a})$ in the worst-case a, where w_a is the proportion of group a. This means that good generalization performance hinges on collecting adequate amounts of data from minority groups. Lastly, as discussed in Section 3.3—although not empirically explored here—higher accuracies could be achieved if the scores were calibrated prior to post-processing.

6. Further Related Work

Fairness Criteria. This paper focused on the *group* criterion of demographic parity, which is defined on population-level statistics (Verma & Rubin, 2018; Kearns et al., 2018). Other group criteria include parity of true positive and/or negative rates (Hardt et al., 2016), predictive rates (Chouldechova, 2017; Berk et al., 2021; Zeng et al., 2022b), accuracy (Buolamwini & Gebru, 2018), etc. Besides group-level, there are criteria defined on the *individual-level* (Dwork et al., 2012; Kearns et al., 2019), which require the model to output similar predictions to individuals deemed to be similar under application and context-specific measures designed by the practitioner.

Mitigation Methods. In addition to post-processing, there are data pre-processing methods (Calmon et al., 2017), as well as in-processing ones via constrained optimization (Kamishima et al., 2012; Zafar et al., 2017; Agarwal et al., 2019) or fair representation learning (Zemel et al., 2013; Madras et al., 2018; Zhao et al., 2020; Song et al., 2019). There are methods under other learning settings and 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 an effective post-processing method with suboptimality and sample complexity analyses. Our implementation uses linear program 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 time efficient optimization methods, e.g., gradient descent (Staib et al., 2017).

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 which is neither strictly convex nor superlinear. 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. The work of HZ was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Cooperative Agreement Number: HR00112320012, a Facebook Research Award, and Amazon AWS Cloud Credit.

References

Abid, A., Farooqi, M., and Zou, J. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 298–306, 2021.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair Regression: Quantitative Definitions and Reduction-Based Algorithms. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 120–129, 2019.

Alghamdi, W., Hsu, H., Jeong, H., Wang, H., Michalak, P. W., Asoodeh, S., and Calmon, F. P. Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection. In *Advances in Neural Information Processing Systems*, 2022.

Ambrosio, L. and Pratelli, A. Existence and stability results in the L^1 theory of optimal transportation. In *Optimal Transportation and Applications*, volume 1813 of *Lecture Notes in Mathematics*, pp. 123–160. Springer, 2003.

Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., and Wagner, T. Scalable Fair Clustering. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 405–413, 2019.

Barocas, S. and Selbst, A. D. Big Data's Disparate Impact. *California Law Review*, 104(3):671–732, 2016.

Barocas, S., Hardt, M., and Narayanan, A. *Fairness and Machine Learning: Limitations and Opportunities*. The MIT Press, 2023.

Bauer, U., Kerber, M., Roll, F., and Rolle, A. A unified view on the functorial nerve theorem and its variations. *Expositiones Mathematicae*, 2023.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, 50 (1):3–44, 2021.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, 2016.

Buolamwini, J. and Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, pp. 77–91, 2018.

Calders, T., Kamiran, F., and Pechenizkiy, M. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18, 2009.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized Pre-Processing for Discrimination Prevention. In *Advances in Neural Information Processing Systems*, 2017.

Caton, S. and Haas, C. Fairness in Machine Learning: A Survey, 2020. *arXiv:2010.04053 [cs.LG]*.

Chen, Y., Telgarsky, M., Zhang, C., Bailey, B., Hsu, D., and Peng, J. A Gradual, Semi-Discrete Approach to Generative Network Training via Explicit Wasserstein Minimization. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 1071–1080, 2019.

Chi, J., Shen, J., Dai, X., Zhang, W., Tian, Y., and Zhao, H. Towards Return Parity in Markov Decision Processes. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 1161–1178, 2022.

Chierichetti, F., Kumar, R., Lattanzi, S., and Vassilvitskii, S. Fair Clustering Through Fairlets. In *Advances in Neural Information Processing Systems*, 2017.

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

Chzhen, E. and Schreuder, N. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *Advances in Neural Information Processing Systems*, 2019.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair Regression with Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, 2020.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 797–806, 2017.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., and Kalai, A. T. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency*, pp. 120–128, 2019.

Denis, C., Elie, R., Hebiri, M., and Hu, F. Fairness guarantee in multi-class classification, 2023. *arxiv:2109.13642 [math.ST]*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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, pp. 4171–4186, 2019.

Diamond, S. and Boyd, S. CVXPY: A Python-Embedded Modeling Language for Convex Optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring Adult: New Datasets for Fair Machine Learning. In *Advances in Neural Information Processing Systems*, 2021.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.

Fish, B., Kun, J., and Lelkes, Á. D. A Confidence-Based Approach for Balancing Fairness and Accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152, 2016.

Forrest, J., Ralphs, T., Santos, H. G., Vigerske, S., Forrest, J., Hafer, L., Kristjansson, B., `jpfasano`, EdwinStraver, Lubin, M., Jan-Willem, `rlougee`, `jpgoncal1`, Brito, S., `h-i-gassmann`, Cristina, Saltzman, M., `tosttost`, Pitrus, B., Matsushima, F., and `to-st`. coin-or/cbc: Release releases/2.10.10, April 2023. URL https://doi.org/10.5281/zenodo.7843975.

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

Gaucher, S., Schreuder, N., and Chzhen, E. Fair learning with Wasserstein barycenters for non-decomposable performance measures. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 2436–2459, 2023.

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. Stochastic Optimization for Large-scale Optimal Transport. In *Advances in Neural Information Processing Systems*, 2016.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online Learning with an Unknown Fairness Metric. In *Advances in Neural Information Processing Systems*, 2018.

Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, 2016.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. Wasserstein Fair Classification. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, pp. 862–872, 2020.

Joseph, M., Kearns, M., Morgenstern, J., and Roth, A. Fairness in Learning: Classic and Contextual Bandits. In *Advances in Neural Information Processing Systems*, 2016.

Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. Fair Algorithms for Infinite and Contextual Bandits, 2017. *arXiv:1610.09559 [cs.LG]*.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-Aware Classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pp. 35–50, 2012.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2564–2572, 2018.

Kearns, M., Roth, A., and Sharifi-Malvajerdi, S. Average Individual Fairness: Algorithms, Generalization and Experiments. In *Advances in Neural Information Processing Systems*, 2019.

Kohavi, R. 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*, pp. 202–207, 1996.

Kull, M. and Flach, P. Novel Decompositions of Proper Scoring Rules for Classification: Score Adjustment as Precursor to Calibration. In *Machine Learning and Knowledge Discovery in Databases*, pp. 68–85, 2015.

Le Gouic, T., Loubes, J.-M., and Rigollet, P. Projection to Fairness in Statistical Learning, 2020. *arXiv:2005.11720 [cs.LG]*.

Li, P., Zhao, H., and Liu, H. Deep Fair Clustering for Visual Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9067–9076, 2020.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 3384–3393, 2018.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*, pp. 107–118, 2018.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of Machine Learning*. The MIT Press, second edition, 2018.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

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

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, 2017.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, 2020.

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

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

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning Controllable Fair Representations. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173, 2019.

Spanier, E. H. *Algebraic Topology*. Springer, 1981.

Staib, M., Claici, S., Solomon, J., and Jegelka, S. Parallel Streaming Wasserstein Barycenters. In *Advances in Neural Information Processing Systems*, 2017.

Vaidya, P. M. Speeding-up linear programming using fast matrix multiplication. In *30th Annual Symposium on Foundations of Computer Science*, pp. 332–337, 1989.

Verma, S. and Rubin, J. Fairness Definitions Explained. In *2018 IEEE/ACM International Workshop on Software Fairness*, 2018.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

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

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. Inequalities for the L_1 Deviation of the Empirical Distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Laboratories, 2003.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N. Learning Non-Discriminatory Predictors. In *Proceedings of the 2017 Conference on Learning Theory*, pp. 1920–1953, 2017.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 962–970, 2017.

Zehlike, M., Bonchi, F., Castillo, C., Hajian, S., Megahed, M., and Baeza-Yates, R. FA*IR: A Fair Top-k Ranking Algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 1569–1578, 2017.

Zemel, R., Wu, Y. L., Swersky, K., Pitassi, T., and Dwork, C. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333, 2013.

Zeng, X., Dobriban, E., and Cheng, G. Bayes-Optimal Classifiers under Group Fairness, 2022a. *arXiv:2202.09724 [stat.ML]*.

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

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

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. Conditional Learning of Fair Representations. In *International Conference on Learning Representations*, 2020.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. 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, pp. 15–20, 2018.

A. Additional Discussions in Section 3.1

This section discusses the reduction of Theorem 3.2 to TV-barycenter in the noiseless setting, and the examples in Section 3.1.

A.1. Reduction to TV-Barycenter

When the classification problem μ is noiseless, i.e., the (unconstrained) Bayes error rate is zero, $\min_h \operatorname{err}(h) = 0$, we show that Theorem 3.2 reduces to a relaxed TV-barycenter problem. For strict DP ($\alpha = 0$), this result is previously established by Zhao & 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 a deterministic labeling function $y : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ s.t. $Y = y(X, A)$ (almost surely). Therefore, the Bayes optimal score $f^* = y$, and we have

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

Theorem A.1 (Minimum Fair Error Rate, Noiseless Setting)**.** *Let $\alpha \in [0, 1]$, and suppose μ is noiseless. We have*

$$\min_{h:\Delta_{\mathrm{DP}}(h)\le\alpha} \operatorname{err}(h) = \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'}\|q_a-q_{a'}\|_\infty\le\alpha}} \sum_{a\in[m]} \frac{w_a}{2} W_1(r_a^*, q_a) = \min_{\substack{q_1,\cdots,q_m\in\mathcal{Q}_k \\ \max_{a,a'}\|q_a-q_{a'}\|_\infty\le\alpha}} \sum_{a\in[m]} \frac{w_a}{2} \|p_a - q_a\|_1.$$

This is due to $\operatorname{supp}(p_a) \subseteq \{e_1, \cdots, e_k\}$ sharing the same finite support with any $q_a \in \mathcal{Q}_k$, whereby $\frac{1}{2} W_1(r_a^*, q_a) = \frac{1}{2} W_1(p_a, q_a) = \frac{1}{2}\|p_a - q_a\|_1$. Specifically, recall the fact that W_1 under the 0-1 distance is equal to $\|\cdot\|_1$:

Proposition A.2. *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. We have $W_1(p, q) = \frac{1}{2}\|p - q\|_1$.*

Proof. By definition, under the metric $d(x, y) = \mathbb{1}[x \neq y]$,

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

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

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

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

Theorem A.3 (Optimal Fair Classifier, Noiseless Setting)**.** *Let $\alpha \in [0, 1]$, suppose μ is noiseless, let $y : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ be the ground-truth labeling function, and $(q_1^*, \cdots, q_m^*)$ a minimizer of Equation (4). Define*

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

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

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

We have

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

Proof. We first verify that $(\mathcal{T}_a \circ y_a)\sharp\mu_a^X = q_a^*$, which would imply $\Delta_{\mathrm{DP}}(h) \leq \alpha$ by the constraints in Equation (4). 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 $\sum_i p_a(e_i) - q_a^*(e_i) = 0 \implies \sum_i \max(0, p_a(e_i) - q_a^*(e_i)) = \sum_i \max(0, q_a^*(e_i) - p_a(e_i))$, 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

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

We conclude by combining the error on all groups and invoking Theorem A.1. □

A.2. Examples Regarding the Tradeoff of DP Fairness

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 $r_a^* := f_a^*\sharp\mu_a^X$ 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.4. 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 & Gordon, 2022), it is concluded that in the noiseless setting, the inherent tradeoff is zero if and only if the class prior distributions are the same, $p_a(e_i) = \mathbb{P}_\mu(Y = e_i \mid A = a)$ and $r_a^* = p_a \coloneqq f_a^* \sharp \mu_a^X$ 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.5. Consider the two-group binary classification problem of two inputs, $\mathcal{X} = \{1, 2\}$, given by

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

Note that the class prior 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. Proofs for Section 3

To make our arguments rigorous, we begin by providing 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 $\mathbb{1}[P] = 1$ if the predicate P is true, else 0.

We provide the proofs to Lemma 3.1 and Theorems 3.2 to 3.4. 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 $f^* : \mathcal{X} \to \Delta_k$ be the Bayes optimal score function. The error rate of any randomized classifier $h : \mathcal{X} \to \mathcal{Y}$ can be written as*

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

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

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

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,

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

where the second equality is due to an identity stated in Equation (8). □

Lemma B.5 (Full Version of Lemma 3.1)**.** *Let $f^* : \mathcal{X} \to \Delta_k$ be the Bayes optimal score function, define $r^* := f^* \sharp \mu^X$, and fix $q \in \mathcal{Q}_k$. 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*

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

where $f^{*-1}(s) := \{x \in \mathcal{X} : f^*(x) = s\}$, satisfies

$$\mathrm{err}(h) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma(s, y). \tag{6}$$

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 Equation (6).

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,

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

as desired, where line 3 is due to $\mathbb{P}(f^*(X)=s, h(X)=y, X=x) = \mathbb{1}[f^*(x)=s]\,\mathbb{P}(h(X)=y, X=x)$ 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 Equation (6) 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

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

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 3.2. 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 $q_1, \cdots, q_m \in \mathcal{Q}_k$,

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

Incorporating the α-DP constraint, we get

$$\min_{h: \Delta_{\mathrm{DP}}(h) \le \alpha} \operatorname{err}(h) = \min_{\substack{q_1, \cdots, q_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q_a - q_{a'}\|_\infty \le \alpha}} \min_{h: h_a \sharp \mu_a^X = q_a, \forall a} \operatorname{err}(h) = \min_{\substack{q_1, \cdots, q_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q_a - q_{a'}\|_\infty \le \alpha}} \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a^*, q_a). \qquad \square$$

Proof of Theorem 3.3. By construction, the Markov kernel of the randomized optimal fair classifier $\bar{h}^*(x, a) := \mathcal{T}^*_{r_a^* \to q_a^*} \circ f_a^*(x)$ 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 α-DP because the q_a^*'s satisfy the constraint in Equation (4), and its error rate achieves the minimum in Theorem 3.2.

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

$$\begin{aligned}
\mathbb{P}(\bar{h}^*(X, A) = y \mid A = a) &= \int_{\mathcal{X}} \mathbb{P}(\bar{h}^*(x, a) = y)\, \mathbb{P}_\mu(X = x \mid A = a)\, \mathrm{d}x \\
&= \int_{\mathcal{X}} \mathcal{K}((x, a), y)\, \mathrm{d}\mu_a^X(x) \\
&= \int_{\mathcal{X}} \frac{\gamma_a^*(f_a^*(x), y)}{\gamma_a^*(f_a^*(x), \mathcal{Y})}\, \mathrm{d}\mu_a^X(x) \\
&= \int_{\Delta_k} \frac{\gamma_a^*(s, y)}{\gamma_a^*(s, \mathcal{Y})} \left(\int_{f_a^{*-1}(s)} \mathrm{d}\mu_a^X(x) \right) \mathrm{d}s \\
&= \int_{\Delta_k} \frac{\gamma_a^*(s, y)}{\gamma_a^*(s, \mathcal{Y})} r_a^*(s)\, \mathrm{d}s \\
&= \int_{\Delta_k} \gamma_a^*(s, y)\, \mathrm{d}s = \gamma_a^*(\Delta_k, y) = q_a^*(y),
\end{aligned}$$

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

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

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

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

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

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

Proof of Theorem 3.4. Recall that the α-fair classifier $\bar{h}$ returned from Algorithm 1 is

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

where $(q_1, \cdots, q_m)$ is a minimizer of Equation (4) on $r_a \coloneqq f_a \sharp \mu_a^X$ and α, and $\mathcal{T}^*_{r_a \to q_a}$ is the optimal transport from r_a to q_a.

Denote the L^1 difference between f and f^* by

$$\mathcal{E} \coloneqq \mathbb{E}_{\mu^X}[\|f_A(X) - f_A^*(X)\|_1].$$

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

$$\begin{aligned}
\operatorname{err}(\bar{h}) &= \frac{1}{2} \sum_{a \in [m]} w_a \, \mathbb{E}\big[\|\mathcal{T}^*_{r_a \to q_a} \circ f_a(X) - f_a^*(X)\|_1 \mid A = a\big] \\
&\leq \frac{1}{2} \sum_{a \in [m]} w_a \big(\mathbb{E}[\|\mathcal{T}^*_{r_a \to q_a} \circ f_a(X) - f_a(X)\|_1 \mid A = a] + \mathbb{E}[\|f_a(X) - f_a^*(X)\|_1 \mid A = a]\big) \\
&= \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a, q_a) + \frac{\mathcal{E}}{2} \\
&= \min_{\substack{q'_1, \cdots, q'_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q'_a - q'_{a'}\|_\infty \leq \alpha}} \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a, q'_a) + \frac{\mathcal{E}}{2},
\end{aligned}$$

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 is because $(q_1, \cdots, q_m)$ is a minimizer. Let $(q_1^*, \cdots, q_m^*)$ denote a minimizer of Equation (4) on the distributions of Bayes scores $r_a^* \coloneqq f_a^* \sharp \mu_a^X$ and α, then by Theorem 3.2,

$$\begin{aligned}
\operatorname{err}(\bar{h}) - \operatorname{err}^*_\alpha &\leq \frac{1}{2} \left(\min_{\substack{q'_1, \cdots, q'_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q'_a - q'_{a'}\|_\infty \leq \alpha}} \sum_{a \in [m]} w_a \, W_1(r_a, q'_a) - \sum_{a \in [m]} w_a \, W_1(r_a^*, q_a^*) \right) + \frac{\mathcal{E}}{2} \\
&\leq \frac{1}{2} \left(\sum_{a \in [m]} w_a \, W_1(r_a, q_a^*) - \sum_{a \in [m]} w_a \, W_1(r_a^*, q_a^*) \right) + \frac{\mathcal{E}}{2} \\
&\leq \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a, r_a^*) + \frac{\mathcal{E}}{2} \\
&\leq \frac{1}{2} \sum_{a \in [m]} w_a \, \mathbb{E}[\|f_a(X) - f_a^*(X)\|_1 \mid A = a] + \frac{\mathcal{E}}{2} \\
&= \mathcal{E},
\end{aligned}$$

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 $\int_{\Delta_k \times \Delta_k} \|s - s'\|_1 \, \mathrm{d}\pi_a(s, s') = \int_{\mathcal{X}} \|f_a(x) - f_a^*(x)\|_1 \, \mathrm{d}\mu_a^X(x) = \mathbb{E}[\|f_a(X) - f_a^*(X)\|_1 \mid A = a]$.

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

$$\begin{aligned}
\operatorname{err}(\bar{h}) &= \sum_{a \in [m]} \frac{w_a}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathbb{P}(f_a^*(X) = s, \mathcal{T}^*_{r_a \to q_a} \circ f_a(X) = y \mid A = a) \, \mathrm{d}(s, y) \\
&\geq \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a^*, q_a) \\
&\geq \min_{\substack{q'_1, \cdots, q'_m \in \mathcal{Q}_k \\ \max_{a,a'} \|q'_a - q'_{a'}\|_\infty \leq \alpha}} \sum_{a \in [m]} \frac{w_a}{2} W_1(r_a^*, q'_a) \\
&= \operatorname{err}^*_\alpha,
\end{aligned}$$

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 $W_1(r_a^*, q_a)$. $\square$

C. Proofs for Section 4

We establish the sample complexities in 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 group-wise calibration of the scores can be dropped by adding the error propagation of Theorem 3.4 to the results. The sample complexity for the general case procedure described in Section 4.3 via smoothing 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 follows directly from the concentration of multinoulli random variables (Weissman et al., 2003):

Theorem C.1. *Let $p \in \Delta_d$, $d \geq 2$, and $\hat{p}_n \sim \frac{1}{n}\mathrm{Multinomial}(n, p)$. W.p. at least $1-\delta$, $\|p - \hat{p}_n\|_1 \leq \sqrt{2d\ln(2/\delta)/n}$.*

Corollary 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 $\hat{p}_n \coloneqq \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. W.p. at least $1-\delta$ over the random draw of the samples, $\|p - \hat{p}_n\|_1 \leq \sqrt{2|\mathcal{X}|\ln(2/\delta)/n}$.*

Theorem C.3 (Hoeffding's Inequality)**.** *Let $x_1, \cdots, x_n \in \mathbb{R}$ be i.i.d. random variables s.t. $a_i \leq x_i \leq b_i$ almost surely. W.p. at least $1-\delta$, $|\frac{1}{n}\sum_{i=1}^n (x_i - \mathbb{E}\, x_i)| \leq \sqrt{\sum_{i=1}^n (b_i - a_i)^2/2n^2 \cdot \ln 2/\delta}$.*

Proof of Theorem 4.3. We first bound the error rate, then Δ_{DP}.

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 score due to calibration of f's. Then by Lemma B.4 applied on μ',

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

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 $(q_1, \cdots, q_m)$ denote a minimizer of Equation (4) on the r_a's with α, then

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

where we defined $\hat{w}_a = n_a/n$, line 2 is because $w_a = \Theta(\hat{w}_a)$ for all $a \in [m]$ when $n \geq \Omega(\max_a \ln(m/\delta)/w_a)$, and line 3 is due to $(\hat{q}_1, \cdots, \hat{q}_m)$ being a minimizer of Equation (4) on the $\hat{r}_a$'s. Because $\|s - s'\|_1 \leq 2\ \mathbb{1}[s \neq s']$, by Proposition A.2,

$W_1(\hat{r}_a, r_a) \leq \|\hat{r}_a - r_a\|_1$, so it follows that

$$\begin{aligned}
\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\alpha,f} &\leq O\left(\sum_{a\in[m]} \sum_{s\in\mathcal{R}_a} \hat{w}_a |r_a(s) - \hat{r}_a(s)|\right) \\
&\leq O\left(\sum_{a\in[m]} \hat{w}_a \sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n_a}}\right) \\
&\leq O\left(\max_a \sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}}\right)
\end{aligned}$$

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

Fairness. In the finite case, we can get a stronger result in terms of the ℓ_1-norm. Note that for all $a \in [m]$ and $y \in \mathcal{Y}$,

$$\begin{aligned}
\sum_{y\in\mathcal{Y}} \left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y)\right| &= \sum_{y\in\mathcal{Y}} \left|\sum_{s\in\mathcal{R}_a} r_a(s)\, \mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s) = y) - \sum_{s\in\mathcal{R}_a} \hat{r}_a(s)\, \mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s) = y)\right| \\
&\leq \sum_{y\in\mathcal{Y}} \sum_{s\in\mathcal{R}_a} |r_a(s) - \hat{r}_a(s)|\, \mathbb{P}(\mathcal{T}^*_{\hat{r}_a\to\hat{q}_a}(s) = y) \\
&= \sum_{s\in\mathcal{R}_a} |r_a(s) - \hat{r}_a(s)| \\
&\leq O\left(\sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}}\right)
\end{aligned}$$

w.p. at least $1-\delta$ from applications of Corollary 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 $\hat{q}_a$. This ℓ_1-norm bound directly implies an ℓ_∞-norm bound:

$$\max_{y\in\mathcal{Y}} \left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y)\right| \leq \sum_{y\in\mathcal{Y}} \left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y)\right| = O\left(\sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}}\right).$$

Lastly, because of the constraint in Equation (4) that $\max_{a,a'\in[m]} \|\hat{q}_a - \hat{q}_{a'}\|_\infty \leq \alpha$,

$$\begin{aligned}
\Delta_{\mathrm{DP}}(\hat{h}) &= \max_{\substack{a,a'\in[m] \\ y\in\mathcal{Y}}} \left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \mathbb{P}(\hat{h}(X,a') = y \mid A = a')\right| \\
&\leq \max_{\substack{a,a'\in[m] \\ y\in\mathcal{Y}}} |\hat{q}_a(y) - \hat{q}_{a'}(y)| \\
&\qquad + \max_{\substack{a,a'\in[m] \\ y\in\mathcal{Y}}} \left(\left|\mathbb{P}(\hat{h}(X,A) = y \mid A = a) - \hat{q}_a(y)\right| + \left|\mathbb{P}(\hat{h}(X,a') = y \mid A = a') - \hat{q}_{a'}(y)\right|\right) \\
&\leq \alpha + \max_{a,a'\in[m]} O\left(\sqrt{\frac{|\mathcal{R}_a| \ln(m/\delta)}{n w_a}} + \sqrt{\frac{|\mathcal{R}_{a'}| \ln(m/\delta)}{n w_{a'}}}\right).
\end{aligned}$$

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

The proof of the sample complexities in the continuous case uses the following uniform convergence results with the VC dimension and the pseudo-dimension as the complexity measure. We omit the proofs, but refer readers to (Shalev-Shwartz & 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 on Lines 6 to 9 of Algorithm 2, (to be) stated in Lemma D.5.

Theorem C.4 (VC Dimension Uniform Convergence)**.** *Let $\mathcal{H}$ be a class of binary functions from $\mathcal{X}$ to $\{0,1\}$, p a distribution over $\mathcal{X} \times \{0,1\}$, of which $(x_1, y_1), \cdots, (x_n, y_n) \sim p$ are i.i.d. samples. W.p. at least $1-\delta$ over the random draw of the samples, $\forall h \in \mathcal{H}$,*

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

for some universal constant c, where d is the VC dimension of $\mathcal{H}$ (Definition D.12 with $k = 2$).

Theorem C.5 (Pseudo-Dimension Uniform Convergence)**.** *Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$, $\ell : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a nonnegative loss function upper bounded by M, p a distribution over $\mathcal{X} \times \mathcal{Y}$, of which $(x_1, y_1), \cdots, (x_n, y_n) \sim p$ are i.i.d. samples. W.p. at least $1-\delta$ over the random draw of the samples, $\forall h \in \mathcal{H}$,*

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

for some universal constant c, where is the pseudo-dimension of $\{(x, y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$ (Definition D.14).

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 & 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 $O(k^2/nw_a)$ term coming from the disagreements between $\mathcal{T}_a$ and $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ on $(x_{a,i})_{i\in[n_a]}$ could potentially be improved to $O(k/nw_a)$ if LP is assumed to return an extremal solution (Peyré & 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.

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 LP$(\hat{r}_1, \cdots, \hat{r}_m, \alpha)$, where $\hat{q}_1, \cdots, \hat{q}_m$ is the minimizer of Equation (4) on the $\hat{r}_a$'s with α. 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 Equation (17)), 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 score due to calibration of f's. Then by Lemma B.4 applied on μ',

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

Define $s_{a,j} \coloneqq f_a(x_{a,j})$. By Theorems C.5 and D.15, and a union bound, we have w.p. at least $1-\delta$, for all $a \in [m]$,

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

Because each $\mathcal{T}_a$ is extracted from the empirical optimal transport $\mathcal{T}^*_{\hat{r}_a \to \hat{q}_a}$ using Lines 6 to 9 of Algorithm 2, by the discussion in Appendix D and Lemma D.5, they both agree on all of $(x_{a,i})_{i\in[n_a]}$ 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 $(x_{a,i})_{i\in[n_a]}$ lie on the same hyperplane almost surely, so the number of disagreements is at most $\binom{k}{2} = k(k-1)/2$, and

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

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 $1-\delta$,

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

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 Equation (4) on the r_a's with α. The existence of this transport in $\mathcal{G}_k$ is due to the problem being semi-discrete and Theorem D.1. Define $q'_a(y) = \frac{1}{n_a}\sum_{j=1}^{n_a}\mathbb{1}[\mathcal{T}^*_{r_a\to q_a}(s_{a,j}) = y]$. It follows that

$$\begin{aligned}
\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\alpha,f} &\le \mathcal{E} + \sum_{a\in[m]}\frac{w_a}{2}(W_1(\hat{r}_a, \hat{q}_a) - W_1(r_a, q_a)) \\
&= \mathcal{E} + \sum_{a\in[m]}\frac{w_a}{2}((W_1(\hat{r}_a, \hat{q}_a) - W_1(\hat{r}_a, q'_a)) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a))) \\
&\le \mathcal{E} + O\left(\sum_{a\in[m]}\frac{\hat{w}_a}{2}((W_1(\hat{r}_a, \hat{q}_a) - W_1(\hat{r}_a, q'_a)) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a)))\right) \\
&\le \mathcal{E} + O\left(\sum_{a\in[m]}\frac{\hat{w}_a}{2}((W_1(\hat{r}_a, q_a) - W_1(\hat{r}_a, q'_a)) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a)))\right) \\
&\le \mathcal{E} + O\left(\sum_{a\in[m]}\frac{\hat{w}_a}{2}(W_1(q_a, q'_a) + (W_1(\hat{r}_a, q'_a) - W_1(r_a, q_a)))\right),
\end{aligned}$$

where we defined $\hat{w}_a = n_a/n$, line 3 is because $w_a = \Theta(\hat{w}_a)$ for all $a \in [m]$ when $n \ge \Omega(\max_a \ln(m/\delta)/w_a)$, line 4 is due to $(\hat{q}_1, \cdots, \hat{q}_m)$ being a minimizer of Equation (4) 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.2, $W_1(q_a, q'_a) = \|q_a - q'_a\|_1$, and w.p. at least $1-\delta$, for all $a \in [m]$,

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

where the last line follows from Theorem C.1, 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 $a \in [m]$,

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

by Theorem C.3, because $\|\mathcal{T}^*_{r_a \to q_a}(s_{a,j}) - s_{a,j}\|_1 \leq 2$.

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

$$\mathrm{err}(\hat{h}) - \mathrm{err}^*_{\alpha,f} \leq O\left(\max_a \left(\sqrt{\frac{k \ln(m/\delta)}{n_a}} + \frac{k^2}{n_a} \right) \right).$$

Fairness. By applying Theorem C.4 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 $a \in [m]$,

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

Now, the decision boundaries of the function $s \mapsto \mathbb{1}[\mathcal{T}_a(s)_i = 1] \in \mathcal{G}_{k,i}$ (defined in Equation (17)) 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 $(x_{a,i})_{i \in [n_a]}$ almost surely, thereby

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

Therefore, we conclude that

$$\begin{aligned} \Delta_{\mathrm{DP}}(\hat{h}) &= \max_{\substack{a, a' \in [m] \\ y \in \mathcal{Y}}} \left| \mathbb{P}(\hat{h}(X, A) = y \mid A = a) - \mathbb{P}(\hat{h}(X, a') = y \mid A = a') \right| \\ &\leq \max_{\substack{a, a' \in [m] \\ y \in \mathcal{Y}}} |\hat{q}_a(y) - \hat{q}_{a'}(y)| \\ &\qquad + \max_{\substack{a, a' \in [m] \\ y \in \mathcal{Y}}} \left(\left| \mathbb{P}(\hat{h}(X, A) = y \mid A = a) - \hat{q}_a(y) \right| + \left| \mathbb{P}(\hat{h}(X, a') = y \mid A = a') - \hat{q}_{a'}(y) \right| \right) \\ &\leq \alpha + O\left(\max_a \left(\sqrt{\frac{k + \ln(mk/\delta)}{n_a}} + \frac{k}{n_a} \right) \right). \end{aligned}$$

The theorem them follows from noting that $n_a = \Theta(n w_a)$ when $n \geq \Omega(\max_a \ln(m/\delta)/w_a)$. □

Proof of Theorem 4.5. Let the $(q_1, \cdots, q_m)$ denote a minimizer of Equation (4) on the r_a's with α, 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

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

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 score due to calibration of f. Then by Lemma B.5 applied on μ', the error rate on group a, denoted by $\mathrm{err}_a(\bar{h}_\rho)$, is

$$\mathrm{err}_a(\bar{h}_\rho) = \frac{1}{2} \int_{\Delta_k \times \mathcal{Y}} \|s - y\|_1 \, \mathrm{d}\gamma'_a(s, y), \tag{7}$$

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

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

whereby

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

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

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

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

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

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

On the other hand, the lower bound follows from Equation (7), where

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

D. Optimal Transport Between Simplex and Vertex Distributions

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

$$\Delta_k := \left\{ x \in \mathbb{R}^k : \sum_{i=1}^k x_i = 1, x_j \geq 0, \forall j \in [k] \right\},$$

and its k vertices are $\{e_1, \cdots, e_k\}$. In this section, we study the optimal transport 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

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

Define vectors

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

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:

$$\|x - e_i\|_1 = 1 - x_i + \sum_{j \neq i} x_j = 1 - 2x_i + \sum_j x_j = 2(1 - x_i), \quad \forall x \in \Delta_k, \, i \in [k] \tag{8}$$

(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 $c(x, y) = \|x - y\|_1$. 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:

$$\mathcal{G}_k := \left\{ x \mapsto e_{\arg\min_{i \in [k]} (\|x - e_i\|_1 - \psi_i)} : \psi \in \mathbb{R}^k \right\} \subset \{e_1, \cdots, e_k\}^{\Delta_k} \tag{9}$$

(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 Figure 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 $z \in \mathbb{R}^k$,*

$$g_z(x) = e_i \quad \textit{if} \quad x_j - x_i \leq z_j - z_i \iff x^\top v_{ij} \leq z^\top v_{ij}, \quad \forall j \neq i \tag{10}$$

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

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

$$z_i = \frac{1}{k} + \frac{1}{2}\left(\frac{1}{k}\sum_{j=1}^{k}\psi_j - \psi_i\right), \quad \forall i \in [k] \tag{11}$$

(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

$$\psi_i = 2(z_1 - z_i) = 2z^\top v_{i1}, \quad \forall i \in [k] \tag{12}$$

(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 Equation (11), by Equation (8),

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

Conversely, let $g_z \in \mathcal{G}'_k$, then for the $g_\psi \in \mathcal{G}_k$ constructed in Equation (12),

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

□

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 over $\mathcal{Y} := \{e_1, \cdots, e_k\}$. We study the geometric properties of the optimal solution to the (Kantorovich) transportation problem between p, q under the ℓ_1 cost,

$$\sup_{\gamma \in \Gamma(p,q)} \int_{\Delta_k \times \mathcal{Y}} \|x - y\|_1 \, \mathrm{d}\gamma(x, y), \tag{13}$$

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

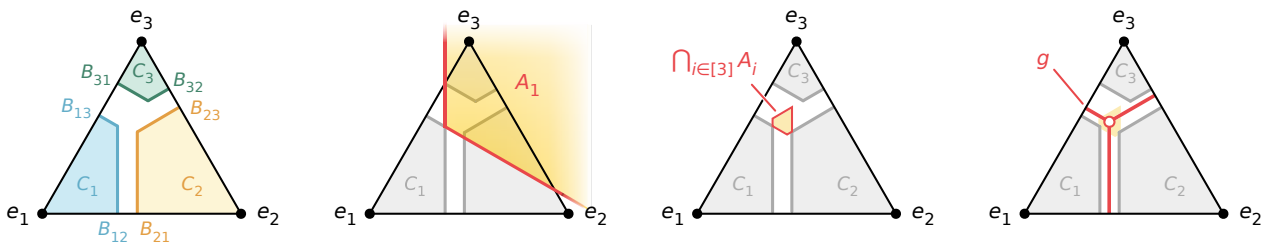

Figure 5. Illustration of the objects defined on Equations (14) and (15) for $k = 3$. See Figure 6 for an example where the intersection is empty, when the underlying transport is not optimal.

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

$$B_{ij} \coloneqq \min\big\{b \in \mathbb{R} : \gamma(\{x \in \Delta_k : x^\top v_{ij} \leq b - 1\}, e_i) = q(e_i)\big\} \cup \{0\}, \quad \text{and}$$
$$C_i \coloneqq \bigcap_{j \neq i} \{x \in \Delta_k : x^\top v_{ij} \leq B_{ij} - 1\}. \tag{14}$$

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 Figures 3 and 5 for illustrations.

Now, if γ^* is an optimal transport of Equation (13), then intuition tells us that in order to achieve minimum cost, the halfspaces along each direction should not overlap (i.e., $B_{ij} + B_{ji} \leq 2$ 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 Figure 5 for an illustration),

$$A_i \coloneqq \bigcap_{j \neq i} \{x \in \mathbb{D}^k : x^\top v_{ij} \geq B_{ij} - 1\}. \tag{15}$$

Proposition D.4. *If γ^* is a minimizer of Equation (13), 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 Equation (13), 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: *(i)* the construction used in its proof led to Lines 6 to 9 of Algorithm 2 for extracting post-processing functions from the empirical optimal transports, *(ii)* it embodies the argument used in the proof of Theorem 4.4 regarding the disagreements between the extracted functions and the empirical transports, and *(iii)* 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 Equation (15) 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 Equation (14). $\square$

D.2. Omitted Proofs for 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

$$d_{ij} = d_{nj} - d_{ni} \quad \text{with} \quad d_{\ell\ell} \coloneqq 0, \quad \forall i, j, n, \ell \in [k], \tag{16}$$

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 Equation (16), $0 < d_{ij} = d_{nj} - d_{ni} < 0$, 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 $i \in [k]$. We have $\bigcap_{i \in [k]} S_i = \emptyset \implies \bigcup_{i \in [k]} S_i \neq \mathbb{D}^k$.*

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., 2023, 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 $i \in [k]$

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

(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

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

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

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

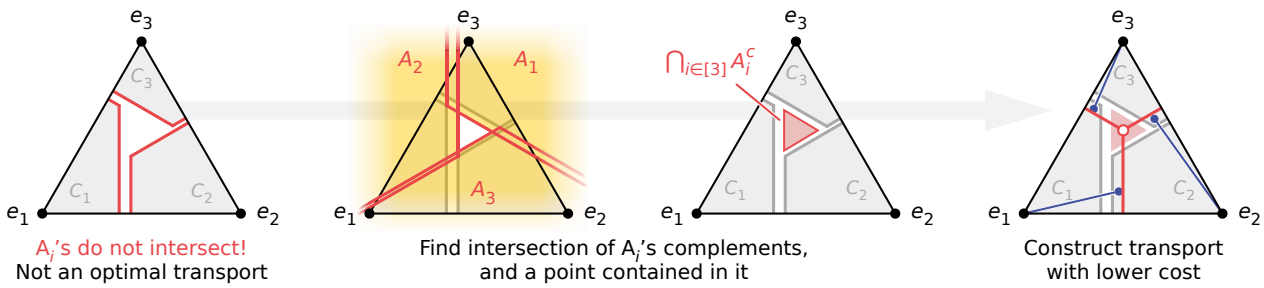

Figure 6. Illustration of the construction in the proof of Proposition D.4 for $k = 3$.

Second, this line is partially not contained any $S'_i = \bigcap_{j\in J}\{x \in \mathbb{D}^k : x^\top v_{ij} \le B_{ij} - 1\}$ 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 a'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 $|J| \le k-1$. 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 $L' := \{z' + a\sum_{i\notin J, j\in J} v_{ij} : a \in \mathbb{R}\}$ is entirely contained in $\bigcap_{i\in J} S'_i$ and partially in $\bigcap_{i\in J} S_i = (\bigcap_{i\in J} S'_i) \cap (\bigcap_{i\notin J, j\in J}\{x \in \mathbb{D}^k : x^\top v_{ij} \le B_{ij} - 1\})$, 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 $\forall J \subset [k]$, $1 \le |J| \le k-1$, $J \in \mathrm{Nrv}(\{S_1, \cdots, S_k\})$. 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 Equations (14) and (15) 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. $\gamma^*(F_i, e_i) := m_i \ge m$.

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

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

has a lower transportation cost than γ^* (see Figure 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 Equation (8). □

Finally, we consider the complexity of the function class $\mathcal{G}_k$ defined in Equation (9). First, recall the definition of multi-class shattering, based on which the Natarajan dimension is defined (Shalev-Shwartz & Ben-David, 2014, Definitions 29.1 and 29.2):

Definition D.11 (Multi-Class Shattering)**.** Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{1,\cdots,k\}$. A set $S \coloneqq \{x_1,\cdots,x_n\} \subseteq \mathcal{X}$ is said to be multi-class shattered by $\mathcal{H}$ if $\exists f_0, f_1 : S \to \{1,\cdots,k\}$ labelings satisfying $f_0(x_i) \neq f_1(x_i)$ for all $i \in [n]$, such that $\forall S_0, S_1$ that partition S (i.e., $S_0 \sqcup S_1 = S$), $\exists h \in \mathcal{H}$, s.t.

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

Definition D.12 (Natarajan Dimension)**.** Let $\mathcal{H}$ be a class of functions from $\mathcal{X}$ to $\{1,\cdots,k\}$. The Natarajan dimension of $\mathcal{H}$, denoted by $d_{\mathrm{N}}(\mathcal{H})$, is the size of the largest subset of $\mathcal{X}$ multi-class 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

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

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

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

Recall from Equation (10) that for all $i, n \in [k]$,

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

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

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

and

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

(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 $j \ne i$, 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 $f_0(x) \ne f_1(x)$ 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 $C_i \coloneqq \bigcap_{j \ne i} \{x \in \Delta_k : x^\top v_{ij} \le B_{ij} - 1\}$. 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

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

(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

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

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 $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. A set $\{x_1, \cdots, x_n\} \subseteq \mathcal{X}$ is said to be pseudo-shattered by $\mathcal{F}$ if $\exists t_1, \cdots, t_n \in \mathbb{R}$ threshold values s.t. $\forall b_1, \cdots, b_n \in \{0, 1\}$ binary labels, $\exists f \in \mathcal{F}$ s.t. $\mathbb{1}[f(x_i) \ge t_i] = b_i$ for all $i \in [n]$.

Definition D.14 (Pseudo-Dimension)**.** Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The pseudo-dimension of $\mathcal{F}$, denoted by $d_{\mathrm{P}}(\mathcal{F})$, is the size of the largest subset of $\mathcal{X}$ pseudo-shattered by $\mathcal{F}$.

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

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. $\mathcal{F}_k$ shatters the set of points. It follows that $\exists g_z, g_{z'} \in \mathcal{G}_k$ s.t. $\|g_{z'}(x_i) - x_i\|_1 < r_i \le \|g_z(x_i) - x_i\|_1$ for all i, which means that $g_z(x_i) \ne g_{z'}(x_i)$. 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 restriction of $\mathcal{G}_k$ to class i by

$$\mathcal{G}_{k,i} \coloneqq \{x \mapsto \mathbb{1}[g(x) = e_i] : g \in \mathcal{G}_k\} \subset \{0,1\}^{\Delta_k}. \tag{17}$$

Because it is a binary function, its *VC dimension*, denoted by $d_{\mathrm{VC}}(\mathcal{G}_{k,i})$, is equivalent to its Natarajan dimension by Definition D.12; moreover, because $\mathcal{G}_{k,i}$ is derived from $\mathcal{G}_k$ by an output remapping, its Natarajan dimension is clearly upper bounded by that of the latter:

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

Adult (Kohavi, 1996)**.** We consider the binary classification task of 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 let gender be the sensitive attribute ($|\mathcal{A}| = 2$). It contains 48,842 examples in total, which we split for pre-training/post-processing/testing by 0.35/0.35/0.3.

ACSIncome (Ding et al., 2021)**.** It is an extension of the Adult dataset with data collected from the US Census Bureau. We consider income prediction under two settings. In the binary setting, the task is to predict whether the annual income of an individual is over or below \$50k per year ($|\mathcal{Y}| = 2$), with gender as the sensitive attribute ($|\mathcal{A}| = 2$). In the multi-group multi-class setting, we create five income buckets of <15000, [15000,30000), [30000,48600), [48600,78030), ≥78030, and group the data into five race categories of "American Indian or Alaska Native alone", "Asian", "Native Hawaiian or Other Pacific Islander alone", "Black or African American alone", "Other", and "White alone" (same as in Adult). It contains 1,664,500 examples, which we split for pre-training/post-processing/testing by 0.63/0.07/0.3.

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 scrapped and hosted by Ravfogel et al. (2020) with 393,423 examples in total, which we split for pre-training/post-processing/testing by 0.35/0.35/0.3.

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).

Communities & Crime (Redmond & 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). It contains 1,994 examples, which we split for pre-training/post-processing/testing by 0.35/0.35/0.3.

E.2. Additional Details

On each task, we first create the pre-training split from the dataset and train a linear logistic regression scoring model using the implementation provided in `scikit-learn` (Pedregosa et al., 2011). Then, we randomly split the remaining data for post-processing and testing with 10 different seeds and aggregate the results as presented in Figures 4 and 7 (the pre-trained model remains the same).

The linear programs of our Algorithm 2 are implemented using the interface of `cvxpy` (Diamond & Boyd, 2016), and are solved using the COIN-OR `Cbc` solver that is based on the *branch and cut* method (Forrest et al., 2023). Finally, the BERT model in BiasBios experiments is loaded through the Hugging Face `Transformers` library (Wolf et al., 2020).

Hyperparameters. The tradeoff curves in Figure 4 and Example A.5 are generated with the following fairness tolerance/strictness settings.

For our method, α is set to:

- ACSIncome (binary). 0.2, 0.18, 0.16, 0.14, 0.12, 0.1, 0.08, 0.06, 0.04, 0.02, 0.01, 0.005, 0.001, 0.0.
- ACSIncome (5-group, 5-class). 0.32, 0.3, 0.28, 0.26, 0.24, 0.22, 0.2, 0.18, 0.16, 0.14, 0.12, 0.1, 0.08, 0.06, 0.04, 0.02, 0.01, 0.0.
- BiasBios. 0.08, 0.07, 0.06, 0.05, 0.04, 0.03, 0.02, 0.01, 0.008, 0.006, 0.004, 0.002, 0.001, 0.0.
- Adult. 0.16, 0.14, 0.12, 0.1, 0.08, 0.06, 0.04, 0.02, 0.01, 0.008, 0.006, 0.004, 0.002, 0.001, 0.0.
- Communities. 0.6, 0.55, 0.5, 0.45, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05, 0.01, 0.0.

For `FairProjection`, we use the default settings that came with the code/package; in particular, increasing the number of iterations to over 1,000 did not improve performance. The tolerance is set to:

- ACSIncome (binary). 0.3, 0.2, 0.18, 0.16, 0.14, 0.12, 0.1, 0.08, 0.06, 0.04, 0.02, 0.01, 0.005, 0.001, 0.0.
- ACSIncome (5-group, 5-class). 0.5, 0.4, 0.35, 0.3, 0.25, 0.2, 0.15, 0.1, 0.08, 0.06, 0.04, 0.02, 0.01, 0.0.
- BiasBios. 1.0, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.05, 0.0.
- Adult. 0.6, 0.5, 0.4, 0.3, 0.2, 0.1, 0.0.
- Communities. 1.0, 0.8, 0.6, 0.4, 0.3, 0.2, 0.1, 0.0.

E.3. Finding Feasible Point on Line 7

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 Figure 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 (of a least-squares problem) that maximizes the margins in our experiments for point-finding:

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

Our preliminary experiments showed that using the quadratic program for point-finding led to better post-processing performance with Algorithm 2 than using the linear program, both in terms of the error rate and Δ_{DP} during inference.

Table 1. Running time (in seconds) of post-processing algorithms under the strictest tolerance setting (see Appendix E.2), averaged over three random splits. Our algorithm is run on a single core of an Intel Xeon Silver 4314 CPU, and `FairProjection` is run on an NVIDIA RTX A6000 GPU.

	ACSIncome		BiasBios	Adult	Communities
Groups	2	5	2	2	4
Classes	2	5	28	2	5
Examples (post-processing split)	116,515		137,698	17095	698
Ours (CPU)	2.56	109	797	0.43	0.38
`FairProjection-KL` (GPU)	33	38	99	15	13

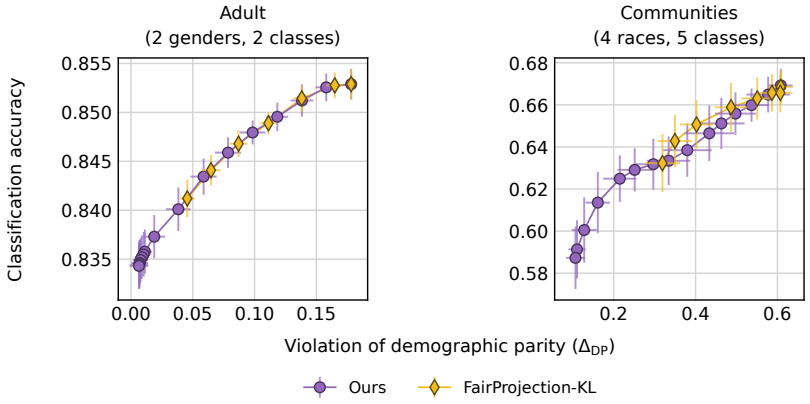

Figure 7. Tradeoff curves between accuracy and Δ_{DP} (Definition 2.1). Scoring model is logistic regression. Error bars indicate the standard deviation over 10 runs with different random splits. Running time is reported in Table 1. On both datasets, `FairProjection-KL` and `FairProjection-CE` have similar results.