Practical Estimation of the Bayes Optimal Fairness-Accuracy Tradeoff with Soft Labels

Mohit Sharma[1*] **Okan Koç**[2] **Amit Deshpande**[3] **Takashi Ishida**[4,2]
Gang Niu[2] **Rajiv Ratn Shah**[5] **Masashi Sugiyama**[4,2]

[1]IIIT Delhi [2]RIKEN-AIP [3]Microsoft Research [4]The University of Tokyo [5]IIT Kanpur

Abstract

There is a fundamental limit to the best prediction error of any model for a given data distribution. In classification, the Bayes error quantifies this limit and serves as both a benchmark for trained models and a criterion to detect overfitting. The state-of-the-art Bayes error estimation techniques for binary classification use only soft labels (positive class probabilities) and require no input features or auxiliary training. We extend such notions to constrained classification problems, such as fairness. Fair classification is an important case of constrained or multi-objective classification. It requires a model not only to be accurate but also to yield fair outcomes (equal or near-equal true positive rates) across different demographic groups (e.g., race and gender). This leads to an inherent fairness-accuracy tradeoff or a Pareto frontier that is used to compare fair classifiers against each other. The focus of our work is to estimate the fundamental limit for this fairness-accuracy tradeoff. We derive an explicit expression for the fair Bayes error rate (i.e., best achievable error rate subject to given fairness constraints), and construct consistent, instance-free estimators for it using only soft labels and group information. We provably bound the bias and the sample complexity of our estimators. Empirically, our method produces a robust, low-variance estimate of the optimal fairness-accuracy tradeoff curve, even when the soft labels are noisy and derived from a classification model on hard labels.

1 Introduction

The notion of Bayes optimality in machine learning defines the best achievable performance, given a data distribution, with respect to the 0-1 loss [1, 2, 3, 4]. For binary classification, an observation consist of features $X \in \mathcal{X} \subseteq \mathbb{R}^d$, a label $Y \in \mathcal{Y} = \{0, 1\}$, and a sensitive attribute $A \in \mathcal{A} = \{0, 1\}$, governed by an underlying distribution $(X, Y, A) \sim \mathcal{D}$. A group-aware classifier $g : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$ predicts labels using both features and sensitive attributes. A well-known result states that the Bayes optimal classifier g^* can be expressed as the following decision rule: $g^*(x, a) = \mathbb{I}\left(\eta(x, a) \geq 1/2\right)$ [4, 5]. Here, $\mathbb{I}\left(\right)$ denotes the indicator function and $\eta(x, a) := \Pr(Y = 1 \mid X = x, A = a)$ defines the posterior, which assigns a soft label $c_i := \eta(x_i, a_i)$ to an instance. This rule also allows us to estimate the Bayes error, the best achievable performance for a given distribution, without access to any instances [4, 6]: $\beta = \mathbb{E}_{X,A}\left[\min\{C, 1 - C\}\right]$, where $C := \eta(X, A)$ is the random variable corresponding to the soft labels.

This notion can also be extended to multi-objective classification problems, such as those that enforce fairness constraints [7]. With the widespread deployment of machine learning models, it is essential that models are fair and do not produce disproportionate errors and harm to certain sensitive groups

*Work done while the author was an intern at RIKEN-AIP. All correspondence to mohit.sharma.cs29@gmail.com.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

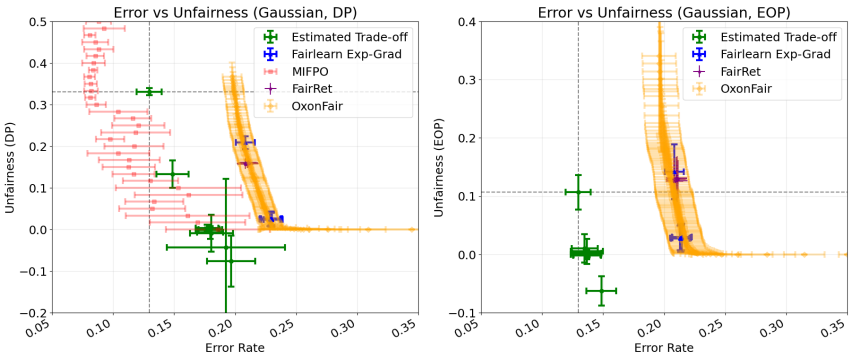


Figure 1: Unfairness-error tradeoffs on synthetic data [12] for Demographic Parity (DP) and Equal Opportunity (EOP) constraints. The proposed estimate (green) establishes a lower bound against competing algorithms. Dotted lines denote the unconstrained Bayes Optimal classifier at $\lambda = 0$.

in the data [7, 8, 9]. However, applying fairness constraints generally introduces a tradeoff with accuracy, since common fairness constraints are incompatible with the standard notion of accuracy [10, 11]. This tradeoff depends sharply on (i) the underlying data distribution, (ii) the fairness metric (equality of positives or true positives), and (iii) the degree of allowed unfairness tolerance.

Bayes optimal fair classifiers can often be characterized as adjustments of the classification thresholds: $g^*_{\text{fair}}(x, a) = \mathbb{I}\left(\eta(x, a) \geq 1/2 + t_a(\delta)\right)$ where $t_a(\delta) > 0$ is a group-dependent adjustment that depends on the fairness metric, the underlying data distribution and the tolerance for unfairness δ. This is precisely where existing works on estimating the Bayes error do not generalize [6, 13, 14]. While Bayes error estimators solely rely on posterior probabilities (soft scores), resulting in a constant threshold of $1/2$, this is not the case for cost-sensitive thresholds, such as fairness. The error estimators for these cost-sensitive thresholds also require an estimator for the underlying thresholds.

Therefore, to estimate the optimal fairness-accuracy tradeoff, we first show how we can disentangle the unfairness level δ and the corresponding threshold adjustment $t_a(\delta)$ by studying when the classifiers become trivial (Lemma 3.8). Using this, we can estimate the threshold and, eventually, the unfairness and error rate at several points along the tradeoff curve by leveraging its monotonicity properties (Lemma C.1). We show that fair classification can be reduced to cost-sensitive thresholding with appropriate λ-dependent thresholds (Lemma C.5), where $\lambda \in \mathbb{R}$ is a parameter that directly controls the correction to the classification threshold and depends on δ. We then derive consistent estimators for the fair classification thresholds, only using soft labels, at any fixed λ (Theorem 4.1).

Once we have consistent estimates of thresholds, we then, as an independent result, generalize the Bayes error estimators in prior works [6, 13, 14] to work with any cost-sensitive risk (Theorem 4.3). Since our thresholds and estimators are both constructed using finite soft labels, we move beyond the scope of prior art to study consistency, bias, and sample complexity of the estimation of the cost-sensitive Bayes risk (Theorem 4.4). These results and rates apply to many cost-sensitive learning scenarios [15, 16] that require distribution-dependent classification thresholds.

We then focus on estimating the optimal fairness-accuracy tradeoff using only soft labels. We first derive the estimated risk at any threshold and show its consistency properties (Theorem 4.5). We then also provide estimators for the expected unfairness under demographic parity and equal opportunity, and derive their consistency and sample complexity properties (Theorem 4.6). All of them culminate in a bisection-search driven algorithm with our estimators (Algorithm 1), inspired by a similar algorithm for the Bayes optimal fair classifiers [5], but only with soft labels and without instances. An example of this bisection algorithm in action is presented in Figure 1 (green curve).

Finally, we estimate the fairness-accuracy tradeoff on real-world datasets. We compare our estimated tradeoff against several recent works on estimating the fairness-accuracy tradeoff [17, 18] and differentiable objectives for fair classification [19], in Figures 1 and 2. We show that our estimated tradeoff lower bounds the tradeoff estimated by these methods. Whenever it does not (for example, in the case of the algorithm proposed in [18]), it is explained by the algorithm's optimization of a completely different lower bound on the Pareto frontier. We also perform experiments with noisy soft

scores generated by a Logistic Regression model on the Adult dataset [20] and find that our method still lower-bounds all other fairness tradeoff algorithms.

Overall, in a similar spirit to prior work [6, 13, 14], where the authors use the Bayes error as a tool to validate performance and check for overfitting, our estimated tradeoff with soft labels establishes a similar precedent: Any current or future methods aiming to compare fair algorithms should always close the gap between their estimated tradeoff and the theoretically best, that we estimate.

Related Work: Our work is inspired by the prior results [6, 13, 14] that characterize the performance limitations in terms of accuracy using the threshold characterization of the Bayes optimal classifier. There are several other works on Bayes error estimation that aim to estimate the Bayes error indirectly [21, 22, 23, 24]. But since most of these methods estimate the Bayes error indirectly, without leveraging soft labels or thresholding rules, generalizing them to constrained optimization settings, such as fairness constraints and cost-sensitive risk, is future work beyond the scope of this work. A major theme of this paper centers around fair machine learning, which has received rightful attention in recent years [7]. Several fairness metrics, such as demographic parity [25] and equal opportunity [26], have been proposed, and corresponding fair Bayes optimal classifiers have been extensively studied in the literature [27, 11]. We make extensive use of threshold characterizations of fair Bayes optimal classifiers [11, 28, 5]. Finally, a major part of our work also concerns deriving consistency and sample-complexity properties of the thresholds and risk and unfairness estimators. Some related estimators have also been studied in prior work [29, 13].

2 Notations and Assumptions

Let $\pi_a = \Pr(A = a)$, $\pi_{ya} = \Pr(Y = y, A = a)$, $\pi_{y|a} = \Pr(Y = y|A = a)$, $\pi = \Pr(Y = 1)$, and $\Pr(Y = y) = \sum_{a \in \mathcal{A}} \pi_{ya}$. We assume the existence of all groups and classes, i.e., $\pi_a > 0, \pi_{y|a} > 0$ for all $y \in \mathcal{Y}$ and for all $a \in \mathcal{A}$. Finally, we let $\xrightarrow{p}$ denote convergence in probability. All probabilities and expectations are taken with respect to the underlying distribution $\mathcal{D}$.

Whenever we work with a finite sample, we assume that we are given access to N i.i.d. soft labels of the form (c_i, a_i), and we will denote our dataset with $D = \{(c_i, a_i)\}_{i=1}^N$. Since our approach is model-free, we do not require access to features x_i. We will also partition our dataset D into two parts: $D = D_\alpha \cup D_\beta$, with $D_\alpha \cap D_\beta = \varnothing$, where D_α (with N_α samples) will be the data used to estimate the thresholds that we will put on $\eta(x, a)$, and D_β (with N_β samples) will be used to construct the estimators of classification risk and unfairness, with $N = N_\alpha + N_\beta$. The splits are chosen independently of the sample values. Furthermore, we will denote with $N_{\alpha,a}$ (or $N_{\beta,a}$) the number of samples from group $A = a$. Lastly, $i \in [D]$ denotes drawing an index from the dataset D. We need two independent sample sets because we estimate the required quantities in two steps. First, using D_α, we estimate the distribution priors and group-dependent thresholds (Theorem 4.1), and next using D_β, we estimate the cost-sensitive Bayes risk, the expected error, and unfairness (Theorems 4.4, 4.5, and 4.6 respectively). We now assume the following about our soft labels:

Assumption 2.1. (No interior atoms / Continuity of the Cumulative Distribution function of η) The mapping $t \to F_a(t) = \Pr(\eta(X, a) \leq t | A = a), \forall a \in \mathcal{A}$ is continuous in $(0, 1)$. Moreover, $\Pr(\eta(X, a) \geq 1/2 | A = a) > 0$ for all $a \in \mathcal{A}$.

Assumption 2.2. (Lipschitz continuous group CDF) For each $a \in \mathcal{A}$, the group-conditional distribution of $C = \eta(X, a)|A = a$ admits a density f_a on $(0, 1)$ such that $\sup_{t \in (0,1)} f_a(t) \leq L_a < \infty$. In particular, F_a is L_a-Lipschitz: $|F_a(u) - F_a(v)| \leq L_a \cdot |u - v|$ for all $u, v \in (0, 1)$.

Let $\hat{\pi}_a = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a)$ denote our estimator for π_a. By the Weak Law of Large Numbers, $\hat{\pi}_a \xrightarrow{p} \pi_a$. Furthermore, the bias of the estimator $\hat{\pi}_a$ is zero, which is easy to see via the linearity of expectation. We can derive a similar unbiased estimator for $\pi = \Pr(Y = 1)$ as $\hat{\pi} = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} c_i$.

Similarly, let $\hat{\pi}_{1a} = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a) c_i$ denote our estimator for π_{1a}. By the Weak Law of Large Numbers, $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$. Furthermore, the bias of the estimator $\hat{\pi}_{1a}$ is zero, which is easy to see via the linearity of expectation. The same analysis holds for the unbiased estimator $\hat{\pi}_{1|a} = \frac{1}{N_{\alpha,a}} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a) c_i$ and its expected value $\pi_{1|a}$. Since we will work with inverse functions of $\hat{\pi}_a$ (and other estimators), we also define $\hat{\pi}_a^+ = \max\{\hat{\pi}_a, \frac{1}{N_\alpha}\}$, which is still statistically consistent,

but a biased estimator (same for $\hat{\pi}^+_{1a}$ and $\hat{\pi}^+_{1|a}$). Finally, let $\Pi_{[0,1]}(t) := \min\{1, \max\{0, t\}\}$ denote the truncation operation, which we will frequently use to ensure our estimators are bounded between 0 and 1. All proofs are provided in the Appendix.

3 Bayes Optimal Fair Classifiers

We will now describe a well-known result that describes the Bayes optimal classifier for binary classification, when we know the underlying distribution.

Lemma 3.1. *[3, 11, 5] The Bayes optimal classifier g^* for a given distribution can be precisely characterized as: $g^*(x, a) = \mathbb{I}(\eta(x, a) \geq 1/2)$.*

However, as noted in the literature [11, 28, 5], the unconstrained Bayes optimal classifier for a distribution may not satisfy fairness constraints. To study this phenomenon in more detail, we now introduce definitions of the fairness constraints we consider in this work.

Definition 3.2. A classifier g is said to be δ-fair if: (1) Demographic Parity (DP) [25]: $|\Pr(g(X, A) = 1|A = 1) - \Pr(g(X, A) = 1|A = 0)| \leq \delta$, and (2) Equal Opportunity (EOP) [26]: $|\Pr(g(X, A) = 1|Y = 1, A = 1) - \Pr(g(X, A) = 1|Y = 1, A = 0)| \leq \delta$.

Intuitively, a classifier satisfies demographic parity whenever it classifies approximately the same number of individuals from different groups as positive. However, this does not account for the error incurred within each group. The equal opportunity constraint overcomes this by enforcing approximately equal true positive rates between the two groups. Zeng et al. [5] recently characterized the Bayes optimal δ-fair classifiers whenever the fairness metric can be expressed as a linear function of a classifier g. We introduce their definition below.

Definition 3.3. A fairness disparity measure $\Delta(g)$ is linear if there exists a $w : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that for all classifiers g, $\Delta(g) = \int_{\mathcal{A}} \int_{\mathcal{X}} g(x, a)\, w(x, a) d\mathbb{P}(x, a)$. Furthermore, it is bilinear if there exist distribution-dependent quantities s, b such that $w(x, a) = s \cdot \eta(x, a) + b$.

Remark 3.4. Proposition 3.4 in Zeng et al. [5] shows the suitable choices of s and b for the fairness constraints mentioned in Definition 3.2. For demographic parity, $w_{DP}(x, a) = \frac{(2a-1)}{\pi_a}$ and for equal opportunity, $w_{EOP}(x, a) = \frac{(2a-1)\eta(x,a)}{\pi_{1a}}$.

We will now define the risk of a classifier g and its disparity for linear disparity metrics [5].

Definition 3.5. (Equation A.1 and 4.2 from [5]) For a classifier g, we can define the risk $R(g)$ and disparity $\Delta(g)$ as: $R(g) = \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} (1 - 2\eta(x, a)) g(x, a) d\mathbb{P}_{X|A=a}(x) + \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} \eta(x, a) d\mathbb{P}_{X|A=a}(x)$, and $\Delta(g) = \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} [w(x, a) \cdot g(x, a)]\, d\mathbb{P}_{X|A=a}(x)$.

Because we are working with binary classification, introducing the class posterior probability allows us to decompose the risk into purely classifier-dependent and distribution-dependent terms. We now introduce the optimal fair classification problem. For at most δ-unfairness, we want to solve the following optimization problem: $g_\delta \in \arg\min_g R(g)$ subject to $|\Delta(g)| \leq \delta$. For demographic parity and equal opportunity, we can precisely write down the fair Bayes optimal classifier that solves this optimization problem, where the threshold depends on the distribution.

Lemma 3.6. *(Proposition 2.3 in [28], Theorem 4.2 in [5]) Under Assumption 2.1, the Bayes optimal δ-fair classifier takes the following form: $g_\delta(x, a) = \mathbb{I}\left(\eta(x, a) \geq \frac{1}{2} + \frac{\lambda^*(\delta)}{2} w(x, a)\right)$, where $\lambda^*(\delta)$ is the optimal dual parameter for the optimization program.*

Remark 3.7. Using Remark 3.4, we can immediately write the threshold characterizations for demographic parity: $g_\delta(x, a) = \mathbb{I}\left(\eta(x, a) \geq \frac{1}{2} + \frac{(2a-1)\lambda^*(\delta)}{2\pi_a}\right)$ and equal opportunity: $g_\delta(x, a) = \mathbb{I}\left(\eta(x, a) \geq \frac{1}{2 - \frac{(2a-1)\lambda^*(\delta)}{\pi_{1a}}}\right)$.

Lemma 3.6 implies that for a fixed δ, there may exist a λ^* (that depends on whether for a given δ, finding a $\lambda^*(\delta)$ is feasible), with which we can adjust the unconstrained Bayes optimal threshold

of $1/2$. This correction depends on which group $A = a$ is underprivileged, as well as on the distribution-dependent quantities like π_a, π_{ya}.

Characterizing the precise value of λ^* as a function of δ and distribution-dependent quantities is a challenging problem, and several prior works (e.g., [11, 30, 28, 5]) treat it as another optimization problem. However, in our approach, we will show how we can obtain a feasible range for λ. By doing so, we will first be able to parametrize the risk and the unfairness in terms of λ, and then use a bisection search to get arbitrarily close to the required level of unfairness δ. In a similar spirit to the Bayes error estimation work in the literature [6, 13, 14], this will allow us to achieve the theoretically optimal fairness-accuracy tradeoff using only soft labels.

Let $g_\lambda = \operatorname{argmin}_g R(g) + \lambda \cdot \Delta(g)$, for $\lambda \in \mathbb{R}$ denote the primal solution for a fixed λ. We dub this as the λ-fair classifier, since this can potentially be a $\delta(\lambda)$-fair classifier. We can now write down the Bayes optimal λ-fair classifier using the threshold characterization in Lemma 3.6, which will be useful later in this paper. We now prove how distribution-dependent quantities bound λ.

Lemma 3.8. *Assuming $w(.,0) < 0$ and $w(.,1) > 0$ (Definition 3.3), when $\Delta(g_0) > 0, \lambda_{\max} = \frac{1}{\inf_{(x,a)} |w(x,a)|}$, so that $\lambda \in [0, \lambda_{\max}]$ is sufficient to interpolate the entire Pareto frontier, including the λ for exact fairness $\lambda^*(0)$. And when $\Delta(g_0) < 0, \lambda_{\min} = \frac{-1}{\inf_{(x,a)} |w(x,a)|}$, so that $\lambda \in [\lambda_{\min}, 0]$, is sufficient to interpolate the entire Pareto frontier.*

This condition becomes much simpler for the equal opportunity and the demographic parity fairness metrics: $\lambda_{\max} = \max_{a \in \mathcal{A}} \pi_{1a}$ and $\lambda_{\max} = \max_{a \in \mathcal{A}} \pi_a$, respectively. Restricting the possible range of λ allows us to set up a bisection search to estimate the fairness-accuracy tradeoff (Algorithm 1). We now argue that approximating this tradeoff curve pointwise is sufficient for our scenario to obtain a high-probability estimate of the true fairness-accuracy tradeoff.

Proposition 3.9. *If a classifier g belongs to the Pareto-frontier, then $g = g_\lambda$, for some $\lambda \in \mathbb{R}$.*

In the Appendix, we show that g_λ can be constructed as a threshold classifier of the form $g_\lambda(x, a) = \mathbb{I}(\eta(x, a) \geq \alpha_{a,\lambda})$, where the threshold $\alpha_{a,\lambda}$ is λ and distribution dependent (Remark 3.7).

4 Estimating the Unfairness and Error Rates Using Soft Labels

With all the tools in place, we will now derive expressions for the expected and finite-sample estimators. We will first derive consistent estimators of the cost-sensitive thresholds α_a that depend on λ, the unfairness metric of choice, and some distribution-dependent quantities in Theorem 4.1. Once we have that, we will first show a consistent estimator for the cost-sensitive Bayes risk, in a similar spirit to the results in prior work [6, 14] in Theorem 4.3 and then its finite sample version in Theorem 4.4. Since we are interested in estimating the fairness-accuracy tradeoff, we will then derive consistent estimators for the risk of a classifier (Theorem 4.5) and then the unfairness (Theorem 4.6). Finally, we will present an algorithm to construct the tradeoff curve (Algorithm 1 in the Appendix).

4.1 Estimating Fair Thresholds using Finite Soft Labels

Unlike prior work [6, 31, 13, 14], where soft labels were enough to estimate the Bayes risk β, in the case of any cost-sensitive risk β_{CS}, unless the cost α_a is a constant, we need to estimate α_a from given soft labels, since α_a would often depend on distribution dependent quantities. This is certainly true for the fairness constraints defined in Lemma 3.6. We now define the estimated thresholds for both DP and EOP fairness constraints (Lemma 3.6 and Remark 3.7). For demographic parity: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2} + \frac{(2a-1)\lambda}{2\hat{\pi}_a^+}$, and for equal opportunity: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2 - \frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}}$, where Π is the truncation operation defined earlier, that ensures that our estimated thresholds stay in the interval $[0, 1]$.

We now show that the estimators of α_a for DP and EOP, with finitely many soft labels, are statistically consistent. Since the specific construction of the estimator requires knowledge of $w(X, A)$, we present estimators for demographic parity and equal opportunity, and the proof techniques can be generalized to tackle other cost-sensitive applications.

Theorem 4.1. *(Consistency and Sample Complexity for threshold estimation) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and a fixed $\lambda \in \mathbb{R}$, consider the empirical estimator $\hat{\alpha}_a^+$. Assuming $2 - \frac{(2a-1)\lambda}{\pi_{1a}} \geq \tau > 0$ (denominator bounded away from zero), the following holds:*

1. *Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.*
2. *Bias:*
   - *DP: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a^+]| \leq \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp(-\frac{\pi_a^2}{2}N_\alpha)$,*
   - *EOP: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a^+]| \leq \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp(-2\epsilon^2 N_\alpha)$, where $\epsilon \leq \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$.*
3. *Sample Complexity: With $1-\delta$ probability, $\Pr(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon) \leq \delta$, as long as:*
   - *DP: $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$,*
   - *EOP: $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.*

Remark 4.2. It is possible to remove the dependency of λ in the sample complexity results by using Lemma 3.8 that tells us about the maximum λ, and study the worst-case sample complexity bounds. However, keeping λ-dependent results here highlights the (indirect) polynomial dependency on the unfairness through the correction factor λ. This also indicates that threshold estimation is non-uniform with respect to the underlying unfairness: the number of samples required may change drastically depending on the degree of correction and the relative group proportions in the distribution.

4.2 Bayes Error Estimation for Cost-Sensitive Classification

Prior works [6, 14, 13] study an estimator for the Bayes error of a distribution for binary classification, by leveraging a simple observation: the Bayes error (which we can denote with β) can be written down as $\beta = \mathbb{E}_X[\min(\eta(x), 1-\eta(x))]$. They also show that $\hat{\beta} = \frac{1}{N}\sum_{i\in[N]}\min(\eta_i, 1-\eta_i)$ is a consistent estimator of the Bayes risk. In the previous section, we derived the expression for cost-sensitive Bayes risk and the corresponding Bayes optimal classifier (Lemma C.5). Let β_{CS} denote the Bayes error for a distribution with respect to a fixed cost-sensitive risk, with group-dependent costs α_a. As an independent result, and in a similar spirit to Proposition 3.1 of Ishida et al. [6], we now derive the consistency and sample-complexity properties of an estimator for the cost-sensitive Bayes risk.

Theorem 4.3. *Consider the Bayes optimal classifier for the cost-sensitive risk with group-dependent costs α_a:*

$$g_\alpha^*(x,a) = \begin{cases} 1, & \text{if } \eta(x,a) \geq \alpha_a \\ 0, & \text{otherwise.} \end{cases}$$

1. *Bayes Risk: The Bayes error with respect to the cost-sensitive risk can be expressed as $\beta_{CS} = \mathbb{E}_{X,A}[\min((1-\alpha_A)\eta(X,A), \alpha_A(1-\eta(X,A)))]$.*
2. *Consistency: $\hat{\beta}_{CS} = \frac{1}{N_\beta}\sum_{i\in[N_\beta]}\min((1-\alpha_i)c_i, \alpha_i(1-c_i)) \xrightarrow{p} \beta_{CS}$.*
3. *Sample Complexity: For any $\delta > 0$, with probability at least $1-\delta$, $|\hat{\beta}_{CS} - \beta_{CS}| \leq \frac{1}{4}\sqrt{\frac{1}{N_\beta}\log(\frac{2}{\delta})}$.*

While this result is geared toward group-aware cost-sensitive risks, it provides a roadmap for deriving results similar to prior work [6, 14, 13] and for generalizing their estimators for Bayes error estimation in cost-sensitive and complex metric applications [15, 16]. We will now proceed to specialize Theorem 4.3 towards fairness applications and construct finite sample estimators using soft labels.

With the fact $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ established in Theorem 4.1, let $\hat{\beta}_{CS,\hat{\alpha}_a^+} = \frac{1}{N_\beta}\sum_{i\in[D_\beta]}\min((1-\hat{\alpha}_a^+)c_i, \hat{\alpha}_a^+(1-c_i))$ denote the empirical estimator of the cost-sensitive Bayes error with the empirical estimation of cost-sensitive thresholds α_a. We now show that $\hat{\beta}_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

Theorem 4.4. *Suppose the dataset D_α with N_α samples is used to construct a consistent estimator of the cost-sensitive threshold α_a (Theorem 4.1). Then, given a dataset D_β with N_β samples, and*

assuming $2 - \frac{(2a-1)\lambda}{\pi_{1a}} \geq \tau > 0$ *(denominator bounded away from zero), the following statements are true:*

1. *Consistency:* $\hat{\beta}_{CS,\hat{\alpha}} = \frac{1}{N_\beta} \sum_{i \in [D_\beta]} \min((1-\hat{\alpha}^+_{a_i})c_i, \hat{\alpha}^+_{a_i}(1-c_i)) \xrightarrow{p} \beta_{CS}$.
2. *Bias: For a fixed* $\delta \in (0,1)$*:*
   - *Demographic Parity:* $|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(|\lambda| \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,
   - *Equal Opportunity:* $|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(\frac{|\lambda|}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.
3. *Sample Complexity: With* $1-\delta$ *probability,* $|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| < \epsilon$ *as long as* $N_\beta \geq \frac{1}{8\epsilon^2}\log(\frac{4}{\delta})$ *and:*
   - *Demographic Parity:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
   - *Equal Opportunity:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\tau^4\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.

Unlike Bayes error estimation [6, 13, 14] where $\alpha_a = 1/2$, due to distribution and unfairness dependent α_a, we now have biased estimators and sufficient sample requirement that depends on the unfairness correction λ, and for τ that determines how well our denominator is defined.

4.3 Expected Error Estimation

However, when considering the fairness-accuracy tradeoff, we are interested in computing the error rate with respect to the expected 0-1 loss. Recall from Definition 3.5 the definition of expected risk with respect to the 0-1 loss: $R(g) = \sum_{a \in \{0,1\}} \pi_a \Big(\mathbb{E}_{X|A=a}\left[(1-2\eta(X,a))g(X,a)\right] \Big) + \pi = \mathbb{E}_{X,A}\left[(1-2C)\mathbb{I}\left(C \geq \alpha_A\right)\right] + \pi$. A finite sample estimator of the above risk can be written down as follows: $\hat{R}(\hat{g}) = \frac{1}{N_\beta} \sum_{i \in [D_\beta]} (1-2c_i)\mathbb{I}\left(c_i \geq \hat{\alpha}^+_{a_i}\right) + \hat{\pi}$.

Theorem 4.5. *Consider* $\hat{R}(\hat{g})$*, the empirical estimator for the expected risk from Definition 3.5. Then, the following properties hold:*

1. *Consistency:* $\hat{R}(\hat{g}) \xrightarrow{p} R(g)$.
2. *Bias: For a fixed* $\delta \in (0,1)$*:*
   - *Demographic Parity:* $|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(|\lambda| \cdot \max\{L_0, L_1\} \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,
   - *Equal Opportunity:* $|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda|\max\{L_0,L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.
3. *Sample Complexity: With* $1-\delta$ *probability,* $|R(g) - \hat{R}(\hat{g})| < \epsilon$ *as long as* $N_\beta \geq \frac{18}{\epsilon^2}\log\left(\frac{6}{\delta}\right)$ *and:*
   - *Demographic Parity:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
   - *Equal Opportunity:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$.

Note that estimating the expectation using the sample set D_β is standard. The difficult part is that our underlying group-dependent thresholds are estimates with respect to D_α. This makes estimating risk under constraints more difficult and introduces nontrivial differences from prior results on risk estimation with finite samples [6, 13, 14].

4.4 Unfairness Estimation

We will now estimate unfairness using a finite number of soft labels. Since every fairness metric would require different distribution-dependent quantities, we will consider demographic parity and equal opportunity metrics here. To proceed with estimating the unfairness rates, we will need estimators for the True positive rates (TPR) and Positive rates (PR) later in the paper. We will first tackle TPR,

and PR can be tackled similarly. A natural estimator for the TPR is to replace all quantities involved with their sample estimates, as done in Jeong et al. [13]. Let $\rho_{\mathrm{TPR},a}(g_\alpha)$ denote the true positive rate for a threshold classifier $g_\alpha(x,a) = \mathbb{I}(\eta(x,a) \geq \alpha_a)$ on group $A=a$, which can be expressed as: $\rho_{\mathrm{TPR},a}(g_\alpha) = \Pr(g_\alpha(X,A)=1|Y=1,A=a) = \frac{\mathbb{E}_{X|A=a}[\mathbb{I}(\eta(X,a)\geq\alpha_a)\eta(X,a)]}{\pi_{1|a}}$, and its sample estimate: $\hat{\rho}_{\mathrm{TPR},a}(g_{\hat{\alpha}_a^+}) = \frac{\frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}(c_i(\mathbb{I}(c_i\geq\hat{\alpha}_a^+))}{\hat{\pi}_{1|a}^+}$, where $N_{\beta,a}$ are the number of samples in D_β from group $A=a$, and $\hat{\alpha}_a^+$ is the threshold constructed using finite samples.

Once we have our estimators for the TPR, we can define the equal opportunity difference of a classifier g as $\Delta_{\mathrm{EOP}}(g) = |\rho_{\mathrm{TPR},1}(g) - \rho_{\mathrm{TPR},0}(g)|$ and its empirical counterpart $\hat{\Delta}_{\mathrm{EOP}}(g) = |\hat{\rho}_{\mathrm{TPR},1}(g) - \hat{\rho}_{\mathrm{TPR},0}(g)|$. This defines $\Delta_{\mathrm{EOP}}(g_\alpha)$ as the true equal opportunity difference with respect to the true threshold α_a and $\hat{\Delta}_{\mathrm{EOP}}(g_{\hat{\alpha}})$ as the empirical estimate of equal opportunity with respect to the empirical estimate of the threshold.

If we only care about the positive rate (for demographic parity), we can write: $\rho_{\mathrm{PR},a}(g_\alpha) = \Pr(g_\alpha(X,a)=1|A=a) = \mathbb{E}_{X|A=a}[\mathbb{I}(\eta(X,a)\geq\alpha_a)]$, and its finite sample estimator: $\hat{\rho}_{\mathrm{PR},a}(g_{\hat{\alpha}_a^+}) = \frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}\mathbb{I}(c_i\geq\hat{\alpha}_a^+)$. Similar to EOP, we can write the DP difference as $\Delta_{DP}(g) = |\rho_{\mathrm{PR},1}(g) - \rho_{\mathrm{PR},0}(g)|$. Then, the true DP difference on the true threshold is $\Delta_{DP}(g_\alpha)$ and the empirical DP difference on the estimated threshold is $\hat{\Delta}_{DP}(g_{\hat{\alpha}})$.

Theorem 4.6. *Suppose a dataset D_β is used to construct $\hat{\Delta}(g_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the DP or the EOP threshold (Theorem 4.1). Then the following holds:*

1. *Consistency:* $\hat{\Delta}(g_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta(g_\alpha)$.
2. *Bias: For $\delta \in (0,1)$,*
   - *DP:* $|\Delta_{DP}(g) - \mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]| = \mathcal{O}\left(\frac{1}{\sqrt{N_\beta}} + |\lambda|\cdot\max\{L_0,L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right)$.
   - *EOP:* $|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda|\cdot\max\{L_0,L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}}\right)$.
3. *Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{g}) - \Delta(g)| \leq \epsilon$, as long as $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ and:*
   - *DP:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\cdot\max\{L_0^2,L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
   - *EOP:* $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2,L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$.

Just like the bias and sample complexity results in Theorem 4.5, the rates of unfairness estimation depend sharply on the unfairness correction and our smoothness assumptions, apart from standard rates [4]. For readability, we subsume the dependence on prior group and subgroup probabilities (π_a and π_{ya}) within the order notation. However, the proofs in the Appendix highlight this dependence explicitly, which is useful to study rare and imbalanced group behaviors. We provide a proof sketch, along with the proofs in the Appendix.

Tradeoff Estimation: We use Algorithm 1 (details in the Appendix) to estimate the Pareto frontier using the estimators derived using soft labels in the previous sections. The tradeoff algorithm appeals to the monotonicity and convexity of the Pareto frontier (Lemma C.1 and Proposition 4.6 in [5]) to set up a bisection search. However, since the guarantees for our estimators in previous sections are only derived for a fixed but arbitrary λ, we below note a couple of approaches to derive high-probability guarantees on the estimated Pareto frontier.

Remark 4.7 (On the consistency of the estimated frontier). Our aim is to estimate the fairness-accuracy tradeoff, estimated pointwise in λ, rather than the risk at a prescribed unfairness level δ: the guarantees of Theorem 4.1 (and Theorems 4.5, 4.6) are stated for a fixed λ. However, we can provide a high-probability guarantee on the 'correctness' of the tradeoff: (1) Fixed grid: If the λ values are fixed on a grid $\{\lambda_1,\ldots,\lambda_K\} \subset [\lambda_{\min},\lambda_{\max}]$, the search points are data independent, and a union bound over the K points controls the error; (2) Bisection: In its current state, Algorithm 1 selects each λ from the estimated disparity $\hat{\Delta}$, which makes the points data-dependent, and a union bound over them is not directly valid. This can be rectified by studying the uniform-in-λ guarantee,

$\sup_\lambda |\hat{\Delta}(h_\lambda) - \Delta(h_\lambda)| \to 0$. Since $\{h_\lambda\}$ is a low-complexity, one-parameter monotone threshold family, such a bound may be available via a uniform law of large numbers. Bisection is more query-efficient, but its analysis requires controlling the error uniformly in λ.

Note that Algorithm 1 can not only be used to generate an estimate of the Pareto frontier, but can also be used to approximate the risk of the classifier when we desire disparity at most δ. This is because we can always interpolate the curve between the obtained points to get a sense of the sharpness of the tradeoff, along with the bias and sample-complexity bounds in Theorems 4.5 and 4.6. This can be readily used to enforce fairness and choose between fairness metrics, among many applications.

5 Experiments

We now present empirical verification of the estimated tradeoff. We work with different sources of soft labels: a synthetic Gaussian setting where we can analytically compute the soft labels [12], an average of hard labels on the CIFAR-10H dataset [32], directly borrowed from prior work, [6, 14] and noisy soft labels derived from a logistic regression model trained on the Adult dataset [20], a popular dataset for unfairness evaluation [7]. Further details of the experiments are provided in the Appendix.

We compare our computed tradeoff against several algorithms that return a tradeoff. We choose the 'Exp-Grad' algorithm [33] due to its theoretical and practical strengths [34, 35], Oxonfair [17], Fairret [19], and the MIFPO algorithm [18]. Oxonfair [17] uses validation data to optimize classification thresholds of a pretrained predictor and supports a variety of fairness metrics. MIFPO (Model-Independent Fairness-Performance Optimization) [18] proposes a discrete optimization method to compute the fairness-accuracy tradeoff using soft scores from calibrated classifiers. Finally, Fairret [19] proposes differentiable regularizers as proxies for any linear-fractional form of fairness metrics [30]. All methods return a tradeoff for both fairness metrics, except MIFPO [18], which returns a tradeoff only for DP. For a fair comparison, we ensure that MIFPO uses the same soft scores we do, rather than recomputing them. Note that our estimated tradeoff curve relies on bisection search and hence concentrates more around the unfairness value of 0, which is our desired target unfairness.

In Figures 1 and 2, we show the results for the Gaussian, CIFAR-10H, and the Adult datasets. Averaged across 5 runs, we see that our estimate of the Pareto frontier consistently lower bounds those of all other methods and exhibits the least variance. A peculiar observation is from the demographic parity results, where MIFPO seems to be a lower bound for the method in almost all cases. This is where we would like to highlight how MIFPO operates. Given a dataset, MIFPO aims to optimize for a lower bound that would almost always lower bound any Pareto frontier obtained on the data (Section 4 in [18]). However, for MIFPO, there are no guarantees that we are optimizing for the Bayes optimal fairness-accuracy tradeoff. The MIFPO algorithm does not optimize the true tradeoff curve and instead relies on convex programming to solve for a proxy objective. Hence, comparing algorithms against MIFPO may always yield a large gap, and such comparisons are not indicative of the true gap.

Finally, note that our method at $\lambda = 0$ corresponds to the Bayes error, as indicated by the intersection of the dotted black lines in Figures 1 and 2. Note that our plots also indicate that many fairness tradeoff estimation techniques fail to remain close to the Bayes error when their objectives don't require fairness-constraining solutions. This might indicate that these algorithms are overfitting, similar to the phenomena observed in prior works on Bayes error estimation [6, 14].

6 Conclusion

In this work, we set out to extend Bayes error estimation techniques towards estimating the fairness-accuracy Pareto frontier. We do so by generalizing the instance-free soft-label estimation techniques from prior work [6, 13, 14] for cost-sensitive risks. Since fairness constraints themselves can be treated as a cost-sensitive learning problem, and given the properties of the fairness-accuracy Pareto front, we can set up consistent estimators for the fair thresholds and for risk and unfairness estimates. Empirically, we can also lower-bound many recent algorithms to obtain the fairness-accuracy frontier.

Our guarantees show that, even with clean soft scores, the problem remains challenging due to strong dependence on the distribution's priors and the unfairness-dependent term λ. Recent work generalizes Bayes error estimation with soft labels to the case of monotonic noise in clean soft

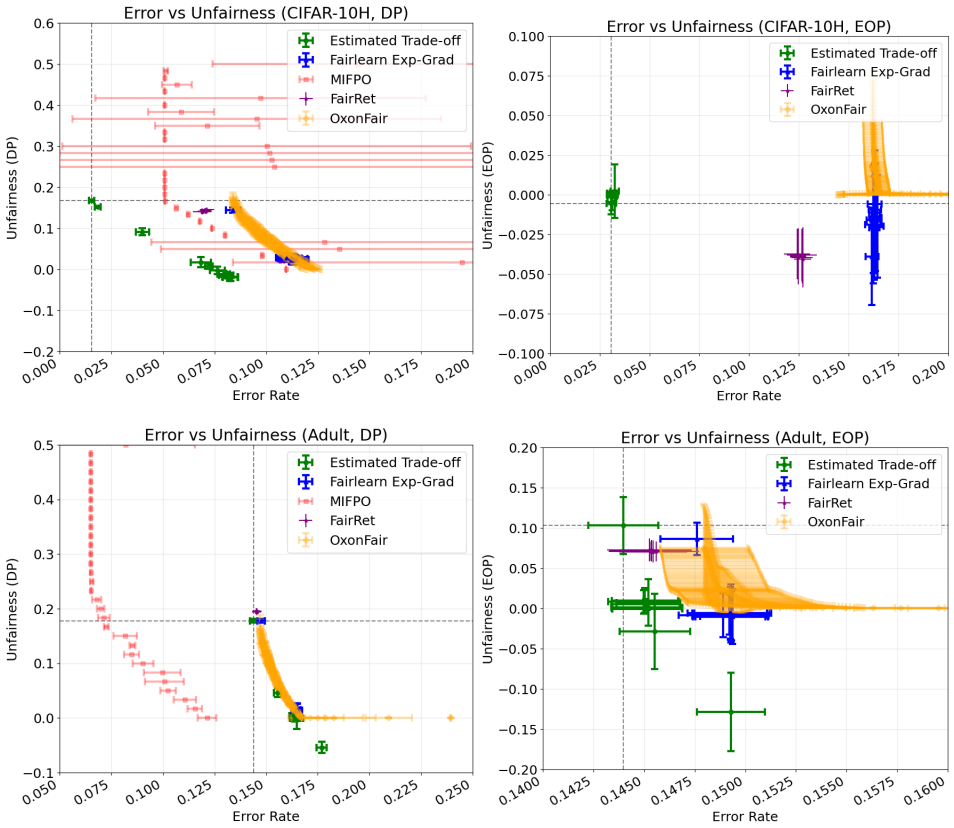


Figure 2: Unfairness-error tradeoffs on Cifar-10H and the Adult Dataset, for Demographic Parity (DP) and Equal Opportunity (EOP) constraints. The proposed estimate (green) establishes a lower bound against competing algorithms, except MIFPO (red), which computes a high-variance data-dependent lower bound different from the Bayes optimal. Dotted lines denote the unconstrained Bayes Optimal classifier at $\lambda = 0$.

scores [14], and extending our estimators to such noisy setups and beyond is an immediate future work. One can also study other models of weak supervision in the Bayes optimal settings, such as positive-confidence (PC) data [36], and positive-unlabeled (PU) data [37]. Generalizing to these settings has attracted attention [38, 6], and fine-tuning our theoretical framework for them is an interesting future direction. To move towards group-blind and sensitive attribute-unaware estimation, we can either use group-blind optimal characterizations (Proposition E.1 in [28]) or probabilistic predictors of sensitive attributes [39, 40, 41, 42], at the cost of increased sample complexity of the underlying estimators. Finally, our guarantees and approach also serve as a template to approach multi-objective problems where the standard Bayes optimal classifier may not align with the optimal predictor for complex metrics [15, 43, 16] and broader applications [44, 45].

Acknowledgments

The authors would like to thank the reviewers for a positive and constructive exchange during the review process. Takashi Ishida was supported by KAKENHI Grant Number 22K17946. Rajiv Ratn Shah would like to thank IIIT Delhi, where he has formerly been an associate professor. Masashi Sugiyama was supported by JST ASPIRE Grant Number JPMJAP25B1.

References

[1] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1):21–27, 1967.

[2] Keinosuke Fukunaga. Introduction to statistical pattern recognition. Elsevier, 2013.

[3] Luc Devroye, László Györfi, and Gábor Lugosi. A probabilistic theory of pattern recognition, volume 31. Springer Science & Business Media, 2013.

[4] M Mohri, A Rostamizadeh, and A Talwalkar. Foundations of machine learning second edition. In Google Books. MIT Press, 2018.

[5] Xianli Zeng, Guang Cheng, and Edgar Dobriban. Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing. arXiv preprint arXiv:2402.02817, 2024.

[6] Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. Is the performance of my deep network too good to be true? a direct approach to estimating the bayes error in binary classification. In The Eleventh International Conference on Learning Representations.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and machine learning: Limitations and opportunities. MIT press, 2023.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, May, 23(2016):139–159, 2016.

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91. PMLR, 2018.

[10] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. Advances in neural information processing systems, 30, 2017.

[11] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In Conference on Fairness, accountability and transparency, pages 107–118. PMLR, 2018.

[12] Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. arXiv preprint arXiv:2202.09724, 2022.

[13] Minoh Jeong, Martina Cardone, and Alex Dytso. Data-driven estimation of the false positive rate of the bayes binary classifier via soft labels. In 2024 IEEE International Symposium on Information Theory (ISIT), pages 368–373. IEEE, 2024.

[14] R. Ushio, T. Ishida, and M. Sugiyama. Practical estimation of the optimal classification error with soft labels and calibration. In Proceedings of Fourteenth International Conference on Learning Representations (ICLR2026), page to appear, Rio de Janeiro, Brazil, Apr. 23–27 2026.

[15] Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent binary classification with generalized performance metrics. Advances in neural information processing systems, 27, 2014.

[16] Shashank Singh and Justin Khim. Optimal binary classification beyond accuracy. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, Advances in Neural Information Processing Systems, 2022.

[17] Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. Oxonfair: A flexible toolkit for algorithmic fairness. Advances in Neural Information Processing Systems, 37:94209–94245, 2024.

[18] Mark Kozdoba, Binyamin Perets, and Shie Mannor. Efficient fairness-performance pareto front computation. arXiv preprint arXiv:2409.17643, 2024.

[19] Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. Fairret: a framework for differentiable fairness regularization terms. arXiv preprint arXiv:2310.17256, 2023.

[20] Moshe Lichman et al. Uci machine learning repository, 2013.

[21] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias. Empirically estimable classification bounds based on a nonparametric divergence measure. IEEE Transactions on Signal Processing, 64(3):580–591, 2015.

[22] Morteza Noshad, Li Xu, and Alfred Hero. Learning to benchmark: Determining best achievable misclassification error from training data. arXiv preprint arXiv:1909.07192, 2019.

[23] Ryan Theisen, Huan Wang, Lav R Varshney, Caiming Xiong, and Richard Socher. Evaluating state-of-the-art classification models against bayes optimality. Advances in Neural Information Processing Systems, 34:9367–9377, 2021.

[24] Cedric Renggli, Luka Rimanic, Nora Hollenstein, and Ce Zhang. Evaluating bayes error estimators on real-world datasets with feebee. arXiv preprint arXiv:2108.13034, 2021.

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science (ITCS) Conference, page 214–226. Association for Computing Machinery, 2012.

[26] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29:3315–3323, 2016.

[27] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In Conference on Learning Theory, pages 1920–1953. PMLR, 2017.

[28] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. Advances in Neural Information Processing Systems, 32, 2019.

[29] Drona Khurana, Srinivasan Ravichandran, Sparsh Jain, and Narayanan Unny Edakunni. Statistical guarantees for fairness aware plug-in algorithms. arXiv preprint arXiv:2107.12783, 2021.

[30] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the conference on fairness, accountability, and transparency, pages 319–328, 2019.

[31] Minoh Jeong, Martina Cardone, and Alex Dytso. Demystifying the optimal performance of multi-class classification. Advances in Neural Information Processing Systems, 36:31638–31664, 2023.

[32] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9617–9626, 2019.

[33] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In International Conference on Machine Learning, pages 60–69. PMLR, 2018.

[34] Mohit Sharma, Amit Deshpande, and Rajiv Ratn Shah. On comparing fair classifiers under data bias. arXiv preprint arXiv:2302.05906, 2023.

[35] Avijit Ghosh, Pablo Kvitca, and Christo Wilson. When fair classification meets noisy protected attributes. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, pages 679–690, 2023.

[36] Takashi Ishida, Gang Niu, and Masashi Sugiyama. Binary classification from positive-confidence data. Advances in neural information processing systems, 31, 2018.

[37] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. Advances in neural information processing systems, 27, 2014.

[38] Ziwei Wu and Jingrui He. Fairness-aware model-agnostic positive and unlabeled learning. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 1698–1708, 2022.

[39] Emily Diana, Wesley Gill, Michael Kearns, Krishnaram Kenthapadi, Aaron Roth, and Saeed Sharifi-Malvajerdi. Multiaccurate proxies for downstream fairness. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pages 1207–1239, 2022.

[40] Flavien Prost, Pranjal Awasthi, Nick Blumm, Aditee Kumthekar, Trevor Potter, Li Wei, Xuezhi Wang, Ed H Chi, Jilin Chen, and Alex Beutel. Measuring model fairness under noisy covariates: A theoretical perspective. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pages 873–883, 2021.

[41] Beepul Bharti, Paul Yi, and Jeremias Sulam. Estimating and controlling for equalized odds via sensitive attribute predictors. Advances in neural information processing systems, 36:37173–37192, 2023.

[42] Beepul Bharti, Mary Versa Clemens-Sewall, Paul H Yi, and Jeremias Sulam. Multiaccuracy and multicalibration via proxy groups. arXiv preprint arXiv:2503.02870, 2025.

[43] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. Journal of Machine Learning Research, 20(172):1–59, 2019.

[44] Peter Súkeník and Christoph Lampert. Generalization in multi-objective machine learning. Neural Computing and Applications, 37(30):24669–24683, 2025.

[45] Tobias Wegel, Geelon So, Junhyung Park, and Fanny Yang. On the sample complexity of semi-supervised multi-objective learning. Advances in Neural Information Processing Systems, 38:128421–128476, 2026.

[46] Hal R. Varian. Two problems in the theory of fairness. Journal of Public Economics, 5(3-4):249–260, None 1976.

[47] Charles Elkan. The foundations of cost-sensitive learning. In International joint conference on artificial intelligence, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[48] Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In Conference on learning theory, pages 489–511. PMLR, 2013.

[49] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. Post-processing for individual fairness. Advances in Neural Information Processing Systems, 34, 2021.

A Notations and Assumptions

Our data domain is defined over $\mathcal{X} \times \mathcal{Y} \times \mathcal{A}$, where $\mathcal{X} \subseteq \mathbb{R}^d, \mathcal{Y} = \{0,1\}$ and $\mathcal{A} = \{0,1\}$. We denote the posterior probability of the positive class, given a sample, as $\eta(x,a) = \Pr(Y=1|X=x, A=a)$. We define the soft label for an instance (x_i, a_i) as posterior probability: $c_i = \eta(x_i, a_i)$, and let $C = \eta(X, A)$ denote a random variable. Let $\pi_a = \Pr(A=a)$, $\pi_{ya} = \Pr(Y=y, A=a)$, and $\pi_{y|a} = \Pr(Y=y|A=a)$. By the law of total probability, $\Pr(Y=y) = \sum_{a\in\mathcal{A}} \pi_{ya}$. We assume the existence of all groups and classes, i.e., $\pi_a > 0, \pi_{y|a} > 0$ for all $y \in \{0,1\}$ and for all $a \in \mathcal{A}$.

Whenever we work with finite sample approximations, we assume that we are given access to N i.i.d. soft labels of the form (c_i, a_i), and we will denote our dataset with $D = \{(c_i, a_i)\}_{i=1}^N$. Since our approach is model-free, we do not require access to instances or features x_i. We will also partition

our dataset D into two parts: $D = D_\alpha \cup D_\beta$, with $D_\alpha \cap D_\beta = \varnothing$, where D_α (with N_α samples) will be the data used to estimate the thresholds, and D_β (with N_β samples) will be used to construct the estimators of the required expectation, with $N = N_\alpha + N_\beta$. The splits are chosen independently of the sample values. Furthermore, we will denote with $N_{\alpha,a}$ (or $N_{\beta,a}$) the number of samples from group $A = a$. We now assume the following about our soft labels:

Assumption A.1. (No interior atoms / Continuity of the CDF of η) The mapping $t \to F_a(t) = \Pr(\eta(X,a) \leq t | A = a), \forall a \in \mathcal{A}$ is continuous in $(0,1)$. Moreover, $\Pr(\eta(X,a) \geq 1/2 | A = a) > 0$ for all $a \in \{0,1\}$.

For later results, we would also need a bounded density assumption on our group-conditioned soft labels.

Assumption A.2. (Lipschitz continuous group CDF) For each $a \in \{0,1\}$, the group-conditional distribution of $C = \eta(X,a)|A = a$ admits a density f_a on $(0,1)$ such that $\sup_{t \in (0,1)} f_a(t) \leq L_a < \infty$. In particular, F_a is L_a-Lipschitz: $|F_a(u) - F_a(v)| \leq L_a \cdot |u - v|$ for all $u, v \in (0,1)$.

Let $\hat{\pi}_a = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a)$ denote our estimator for π_a. By the Weak Law of Large Numbers, $\hat{\pi}_a \xrightarrow{p} \pi_a$. Furthermore, the bias of the estimator $\hat{\pi}_a$ is zero, which is easy to see via the linearity of expectation. We can derive a similar estimator for $\pi = \Pr(Y = 1)$ as $\hat{\pi} = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} c_i$.

Similarly, let $\hat{\pi}_{1a} = \frac{1}{N_\alpha} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a)\, c_i$ denote our estimator for π_{1a}. By the Weak Law of Large Numbers, $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$. Furthermore, the bias of the estimator $\hat{\pi}_{1a}$ is zero, which is easy to see via the linearity of expectation. The same analysis works out for the unbiased estimator $\hat{\pi}_{1|a} = \frac{1}{N_{\alpha,a}} \sum_{i \in [D_\alpha]} \mathbb{I}(a_i = a)\, c_i$ and its expected value $\pi_{1|a}$. Since we will work with functions of $\hat{\pi}_a$ (and other estimators) that may have inverses of $\hat{\pi}_a$, we also define another estimator $\hat{\pi}_a^+ = \max\{\hat{\pi}_a, \frac{1}{N_\alpha}\}$, which is still statistically consistent, but a biased estimator (same for $\hat{\pi}_{1a}^+$ and $\hat{\pi}_{1|a}^+$).

We would also require estimators for the True positive rates (TPR) and Positive rates (PR) later in the paper. We will first tackle TPR, and PR can be tackled similarly. A natural estimator for the TPR is to replace all quantities involved with their sample estimates, as done in **(author?)** [13]. Let $\rho_{\mathrm{TPR},a}(g_\alpha)$ denote the true positive rate for a threshold classifier $g_\alpha(x,a) = \mathbb{I}(\eta(x,a) \geq \alpha_a)$ on group $A = a$.

$$\begin{aligned}\rho_{\mathrm{TPR},a}(g_\alpha) = \Pr(g_\alpha = 1 | Y = 1, A = a) &= \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}(\eta(x,a) \geq \alpha_a) \cdot \Pr(Y = 1 | X = x, A = a)\right]}{\pi_{1|a}} \\ &= \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}(C \geq \alpha_a) \cdot C\right]}{\pi_{1|a}}.\end{aligned}$$

We can now define a sample estimate as:

$$\hat{\rho}_{\mathrm{TPR},a}(g_{\hat{\alpha}_a^+}) = \frac{\frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} (c_i (\mathbb{I}(c_i \geq \hat{\alpha}_a^+))}{\hat{\pi}_{1|a}^+},$$

where $N_{\beta,a}$ are the number of samples in D_β from group $A = a$, and $\hat{\alpha}$ is the threshold constructed using finite samples or intermediate estimators using finite samples (for example, thresholds for fairness constraints derived in Lemma 3.6).

Once we have our estimators for the TPR, we can define the equal opportunity difference of a classifier g as $\Delta_{\mathrm{TPR}}(g) = |\rho_{\mathrm{TPR},1}(g) - \rho_{\mathrm{TPR},0}(g)|$ and its empirical counterpart $\hat{\Delta}_{\mathrm{TPR}}(g) = |\hat{\rho}_{\mathrm{TPR},1}(g) - \hat{\rho}_{\mathrm{TPR},0}(g)|$. This defines $\Delta_{\mathrm{TPR}}(g_\alpha)$ as the true equal opportunity difference with respect to the true threshold α_a and $\hat{\Delta}_{\mathrm{TPR}}(g_{\hat{\alpha}})$ as the empirical estimate of equal opportunity with respect to the empirical estimate of the threshold.

If we only care about the positive rate (for ensuring demographic parity), we can come up with a natural estimator:

$$\rho_{\mathrm{PR},a}(g_\alpha) = \Pr\left(g_\alpha = 1 | A = a\right) = \mathbb{E}_{X|A=a}\left[\mathbb{I}\left(c \geq \alpha_a\right)\right],$$

and its finite sample estimate:

$$\hat{\rho}_{\mathrm{PR},a}(g_{\hat{\alpha}_a^+}) = \frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} \mathbb{I}\left(c_i \geq \hat{\alpha}_a^+\right).$$

Similar to the equal opportunity difference, we can write the demographic parity difference as $\Delta_{DP}(g) = |\rho_{\mathrm{PR},1}(g) - \rho_{\mathrm{PR},0}(g)|$. Then, the true DP difference on the true threshold is $\Delta_{DP}(g_\alpha)$ and the empirical DP difference on the estimated threshold is $\hat{\Delta}_{DP}(g_{\hat{\alpha}})$.

B Helper Results

Lemma B.1. *(Finite sample bound for estimation of π) Let $X_1, X_2, ..., X_N \in [0,1]$ be N i.i.d. random variables with mean π. Let $\hat{\pi} = \frac{1}{N}\sum_{i\in[N]} X_i$. For any $\epsilon > 0$, if $N \geq \frac{1}{2\epsilon^2}\log(\frac{2}{\delta})$, then with $1-\delta$ probability over the choice of the data sample, $\hat{\pi} \in (\pi - \epsilon, \pi + \epsilon)$.*

Proof. The proof follows directly from the application of Hoeffding's inequality[2]. □

Lemma B.2. *(Upper bound on Mean Absolute Error) Consider an estimator $\hat{\pi}$ constructed using N samples for some quantity π (Lemma B.1). If $\hat{\pi}$ is an unbiased estimate of π, then $\mathbb{E}_{D_N}\left[|\hat{\pi} - \pi|\right] \leq \sqrt{Var(\hat{\pi})}$.*

Proof. Let $Z = \hat{\pi} - \pi$. Since $\hat{\pi}$ is an unbiased estimate of π, $\mathbb{E}_{D_N}[Z] = 0$, $Var(Z) = \mathbb{E}_{D_N}\left[Z^2\right] = \mathbb{E}_{D_N}\left[|Z|^2\right] = Var(\hat{\pi})$. Observe that by Jensen's inequality, $\mathbb{E}\left[|Z|^2\right] \geq (\mathbb{E}\left[|Z|\right])^2 \implies \mathbb{E}\left[|Z|\right] \leq \sqrt{\mathbb{E}\left[|Z|^2\right]}$, Hence $\mathbb{E}_{D_N}\left[|\hat{\pi} - \pi|\right] \leq \sqrt{Var(\hat{\pi})}$. □

C Proofs for Section 3

Let $g_\lambda = \underset{g}{\operatorname{argmin}}\ R(g) + \lambda \cdot \Delta(g)$, for $\lambda \in \mathbb{R}$ denote the primal solution for a fixed λ. We can also dub this the λ-fair classifier, since this can potentially be $\delta(\lambda)$-fair classifier. Using this λ, we can also write down the Bayes optimal λ-fair classifier using the threshold characterization in Lemma 3.6, which will be useful later in this paper. We now introduce a key lemma [5] that enables us to study the fairness-accuracy Pareto frontier.

Lemma C.1. *(Proposition 4.1 from [5]) As a function of λ, $\Delta(g_\lambda)$ is monotone non increasing, i.e. $\lambda_1 < \lambda_2 \implies \Delta(g_{\lambda_1}) \geq \Delta(g_{\lambda_2})$. The risk $R(g_\lambda)$ is monotone non-increasing on $(-\infty, 0)$ and monotone non-decreasing on $(0, \infty)$.*

Lemma C.1 is incredibly helpful to devise an algorithm to estimate the tradeoff, since it allows us to parametrize the curve with λ, establish the convexity of the tradeoff curve (Proposition 4.6 in [5]), and hence enables the use of bisection search to get close to any level of unfairness required. We will now attempt to bound the range of λ. Along with this information on endpoints and the monotonicity properties of λ from Lemma C.1, this would allow us to set up a bisection search algorithm.

Lemma C.2. *(Lemma 3.8 in the main text) Assuming $w(.,0) < 0$ and $w(.,1) > 0$ (Definition 3.3), when $Dis(g_0) > 0, \lambda_{\max} = \frac{1}{\inf\limits_{(x,a)} |w(x,a)|}$, so that $\lambda \in [0, \lambda_{\max}]$ is sufficient to interpolate the entire Pareto frontier, including the λ for exact fairness $\lambda^*(0)$. And when $Dis(g_0) < 0, \lambda_{\min} = \frac{-1}{\inf\limits_{(x,a)} |w(x,a)|}$, so that $\lambda \in [\lambda_{\min}, 0]$, is sufficient to interpolate the entire Pareto frontier.*

[2] stanford.edu

Proof. To construct our bounds on λ, we will try to study the case of exact fairness to obtain the worst possible bounds. Assume WLOG that $Dis(g_0) > 0$. The proof for $Dis(g_0) < 0$ is similar.

For demographic parity and equal opportunity fairness constraints (Definition 3.3), $w(.,0) < 0$ and $w(.,1) > 0$. Consider the threshold classifier for groups $A = 0$ and $A = 1$ at any λ (Lemma 3.6):

$$g_\lambda(x,0) = \mathbb{I}\left(\eta(x,0) \geq \frac{1}{2} - \frac{\lambda}{2}|w(x,0)|\right) \quad \text{and} \quad g_\lambda(x,1) = \mathbb{I}\left(\eta(x,1) \geq \frac{1}{2} + \frac{\lambda}{2}|w(x,1)|\right).$$

Also, by assumption, disparity is positive at $\lambda = 0$, where disparity is defined as:

$$Dis(g_\lambda) = \pi_1 \int_{\mathcal{X}} |w(x,1)| \cdot g_\lambda(x,1) d\mathbb{P}_{X|A=1}(x) - \pi_0 \int_{\mathcal{X}} |w(x,0)| \cdot g_\lambda(x,0) d\mathbb{P}_{X|A=0}(x) > 0.$$

From Lemma C.1, we can see that as λ increases, the threshold for group $A = 1$, t_1 increases (since $w(.,1) > 0$) and t_0 decreases (since $w(.,0) < 0$). If $t_1 > 1$, then the $g(x,1) = 0, \forall x$, since $\eta(x,1) \in [0,1]$. Similarly, if $t_0 \leq 0$, $g(x,0) = 1, \forall x$.

At that point, the disparity $Dis(g_\lambda)$ at $\lambda = 0$, which was positive to begin with, becomes negative, and hence we cross over the point of zero unfairness. At such a trivial classification point, $R(g_\lambda) > 0$ and $Dis(g_\lambda) < 0$, which is the minimum possible disparity since $Dis(g) \in [-1,1]$.

It remains to check the value of λ at which the group-aware classifier g becomes trivial for each group. Let such a λ be denoted with $\lambda_{\max}$. Such a $\lambda_{\max}$ occurs at:

$$\begin{aligned}
&\text{(A=1): } \frac{1}{2} + \lambda \frac{|w(x,1)|}{2} > 1 \implies \lambda > \frac{1}{\inf_x |w(x,1)|}, \text{ and} \\
&\text{(A=0): } \frac{1}{2} - \lambda \frac{|w(x,0)|}{2} \leq 0 \implies \lambda \geq \frac{1}{\inf_x |w(x,0)|} \\
&\text{Overall: } \implies \lambda > \frac{1}{\inf_x |w(x,a)|}.
\end{aligned}$$

Now, note that due to Lemma C.1, the tradeoff function parametrized by λ is convex, and by increasing λ towards $\lambda_{\max}$ (the disparity point of -1), we would have also crossed the zero disparity point for some $\lambda \in [0, \lambda_{\max}]$. Hence $\lambda \in [0, \lambda_{\max}]$ whenever the Bayes optimal classifier has positive disparity ($Dis(g_0) > 0$). A similar argument follows for the other case when $Dis(g_0) < 0$. □

This condition becomes much simpler for the equal opportunity and the demographic parity fairness metrics: $\lambda_{\max} = \max_{a \in \mathcal{A}} \pi_{1a}$ and $\lambda_{\max} = \max_{a \in \mathcal{A}} \pi_a$, respectively. Restricting the possible range of λ allows us to set up a bisection search to estimate the fairness-accuracy tradeoff (Algorithm 1). We now argue that approximating this tradeoff curve pointwise is sufficient for our scenario to obtain a high-probability estimate of the true Pareto frontier.

Definition C.3. Given a risk measure $R(\cdot)$ and a fairness disparity measure $\Delta(\cdot)$, a classifier g' Pareto-dominates another classifier g, if $R(g') < R(g)$ and $|\Delta(g')| < |\Delta(g)|$. The Pareto frontier is the set of those classifiers that are not Pareto-dominated by any other classifier.

Using a well-known characterization of the Pareto frontier by [46], we can equivalently write the above set as $g_\lambda = \operatorname{argmin}_g R(g) + \lambda \cdot \Delta(g)$. We now show that any point on the Pareto frontier will be covered by any curve that we generate by varying the values of λ specified by the range obtained from Lemma 3.8.

Proposition C.4. *(Proposition 3.9 in the main text) If a classifier g belongs to the Pareto-frontier, then $g = g_\lambda$, for some $\lambda \in \mathbb{R}$.*

Proof. Suppose g belongs to the Pareto-frontier and WLOG assume $\Delta(g) \geq 0$. We know from Proposition 4.6 in [5] that the Pareto front is convex.

Since g lies on the Pareto-frontier, then $(R(g), \Delta(g))$ must lie on the boundary of the above convex set. Otherwise, if $(R(g), \Delta(g))$ lies in the interior, there exists a small neighborhood around $(R(g), \Delta(g))$ that is contained in this convex set, and would include g' such that $R(g') < R(g)$ and $0 < \Delta(g') < \Delta(g)$. Now consider a supporting tangent to the convex set at $(R(g), \Delta(g))$ defined by a normal vector $(\alpha, \beta) \in \mathbb{R}^2$ and margin/shift $\gamma \in \mathbb{R}$ as follows: $\alpha R(h) + \beta\Delta(h) \geq \gamma$, for all classifiers h, and $\alpha R(g) + \beta\Delta(g) = \gamma$. Let $\lambda = \beta/\alpha$. Then $R(g) + \lambda\Delta(g) = R(g) + \beta/\alpha \cdot \Delta(g) = (\alpha R(g) + \beta\Delta(g))/\alpha = \gamma/\alpha$, whereas for any other h, we have $R(h) + \lambda\Delta(h) = R(h) + \beta/\alpha \cdot \Delta(h) = (\alpha R(h) + \beta\Delta(h))/\alpha \geq \gamma/\alpha$. Therefore, g is a minimizer of $R(h) + \lambda g(h)$ objective over all h, and is itself g_λ for $\lambda = \beta/\alpha$.

□

We will show that any g_λ can be constructed as a threshold classifier, where the threshold is a function of some distribution-dependent parameters and λ. And then using Lemma 3.8, we will be able to estimate the risk and unfairness of several points on the Pareto frontier by varying the value of λ in a fixed range.

Cost sensitive classification: In classification, we generally work with the 0-1 loss. However, there can be scenarios where we need to weigh different errors unequally, such as the cost of missing a disease (it may be much riskier not to catch a disease than to falsely detect it in healthy individuals) and the cost of granting loans to undeserving candidates. This is the setting of cost-sensitive classification [47].

Lemma C.5. *(Extension of [48] to the case of subgroups) Let $\alpha \in [0, 1]$. Define the α-CS risk as weighing the false negatives for a group $A = a$ with $(1 - \alpha_a)$ and the false positives with α_a:*

$$R_\alpha(g) = \sum_{a \in \mathcal{A}} \left((1 - \alpha_a)\pi_{1a} \cdot \mathbb{E}_{X|Y=1,A=a} \left[l_{01}(g(X, a)) \right] + \alpha_a \pi_{0a} \cdot \mathbb{E}_{X|Y=0,A=a} \left[l_{01}(g(X, a)) \right] \right)$$

where $\pi_{ya} = \Pr(Y = y, A = a), y \in \{0, 1\}$ and l_{01} is the 0-1 loss that returns 1 whenever the prediction and the true label mismatch. Then, the cost-sensitive group-dependent Bayes optimal classifier can be written down as:

$$g_\alpha(x, a) = \begin{cases} 1, & \text{if } \eta(x, a) \geq \alpha_a \\ 0, & \text{otherwise} \end{cases}$$

The proof for Lemma C.5 is a simple group-aware extension of Scott et al. [48]. We can immediately see that the thresholds for Bayes optimal fair classifiers from Lemma 3.6 can be written down as group-dependent costs [11]. For example, for demographic parity, $\alpha_a = \frac{1}{2} + \frac{(2a-1)\lambda^*(\delta)}{2\pi_a}$. We will exploit Lemmas 3.8 and C.5 in the coming sections to construct estimators for thresholds, risk, and disparity at arbitrary points up until $\lambda_{\max}$ to get an efficient estimate of the Pareto frontier, only using soft labels.

D Proofs for Section 4

Proof Sketch: We summarize the main ideas behind the proofs of the DP and EOP threshold (Theorem 4.1), and the risk and unfairness estimators (Theorems 4.4, 4.5, 4.6).

- **Consistency**: The consistency proofs begin with the Weak Law of Large Numbers for the empirical priors $\hat{\pi}$, $\hat{\pi}_a$, $\hat{\pi}_{ya}$, and $\hat{\pi}_{y|a}$. The threshold estimators $\hat{\alpha}_a$ are then shown to be consistent by applying the Continuous Mapping Theorem [3] to the corresponding plug-in formulas, together with the truncation operator $\Pi_{[0,1]}$. In the EOP case, we additionally require the relevant denominator to stay bounded away from zero. The later consistency results for risk and unfairness then follow by combining the consistency of $\hat{\alpha}_a$ with continuity of the corresponding functionals, along with standard concentration of empirical averages over the independent sample split.

[3] stat.cmu.edu

- **Bias bounds**: For each estimator, we define a suitable good event $\mathcal{G}$ on which the empirical prior of interest is close to its population counterpart. On $\mathcal{G}$, the estimator error is bounded deterministically in terms of the prior estimation error, using elementary algebraic control of reciprocals whenever needed. Taking expectations then reduces to controlling quantities such as $\mathbb{E}[|\hat{\pi} - \pi|]$, which are bounded using explicit variance calculations and standard inequalities for bounded random variables. On the complement $\mathcal{G}^C$, the estimator error is controlled by the uniform boundedness of the estimator, multiplied by $\Pr(\mathcal{G}^C)$, and $\Pr(\mathcal{G}^C)$ is bounded using Hoeffding's inequality. For the later risk and unfairness results, the threshold estimation error is propagated through the corresponding risk or disparity functional using continuity or Lipschitz bounds under Assumptions 2.1 and 2.2.
- **Sample complexity**: The sample complexity results are obtained by combining concentration bounds for the threshold estimators built from D_α with concentration bounds for the empirical risk and unfairness estimators built from the independent split D_β. The total failure probability is then controlled by a union bound. For the later results, the final sample requirements inherit both the threshold-estimation complexity and the Lipschitz dependence of the downstream risk or disparity estimators on the estimated thresholds.

We now provide proofs for all of the results described in Section 4. For convenience and readability, we will prove the DP and EOP results separately whenever applicable.

Theorem D.1. *(DP results from Theorem 4.1 in the main text) (Consistency and Sample Complexity for DP threshold) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and a fixed $\lambda \in \mathbb{R}$, for the case of Demographic Parity, consider the following estimator of the threshold: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2} + \frac{(2a-1)\lambda}{2\hat{\pi}_a^+}$, $\forall a \in \{0,1\}$. Then the following holds:*

1. *Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.*
2. *Bias: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a^+]| \leq \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp(-\frac{\pi_a^2}{2}N_\alpha)$.*
3. *Sample Complexity: With $1-\delta$ probability and $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$, $\Pr(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon) \leq \delta$.*

Proof. We will prove the results for an arbitrary $a \in \{0,1\}$. Due to Lemma 3.8 and our projection operation Π on the empirical estimator, α_a and $\hat{\alpha}_a^+$ are both restricted to the interval $[0,1]$. Using this, note that in our domain of $\lambda \in [0, \lambda_{\max}]$ or $\lambda \in [-\lambda_{\min}, 0]$, $\alpha_a = \Pi(\alpha_a)$, which is a useful property we exploit throughout the proof.

(1) **Consistency**: By the weak law of large numbers, $\hat{\pi}_a \xrightarrow{p} \pi_a$. Consider $\hat{\pi}_a^+$. Since $\pi_a > 0$ by assumption, we can pick a sufficiently large N_α such that $1/N_\alpha \leq \pi_a/2$. Therefore the event $\{\hat{\pi}_a < 1/N_\alpha\}$ is contained in the event $\{\hat{\pi}_a < \pi_a/2\}$ which is further contained in $\{|\hat{\pi}_a - \pi_a| \geq \pi_a/2\}$. Since $\hat{\pi}_a \xrightarrow{p} \pi_a$, $\Pr(|\hat{\pi}_a - \pi_a| \geq \pi_a/2) \to 0$ and hence $\Pr(\hat{\pi}_a^+ \neq \hat{\pi}_a) \to 0$, and hence $\hat{\pi}_a^+ \xrightarrow{p} \pi_a$.

The estimator $\hat{\alpha}_a^+$ is a continuous function of $\hat{\pi}_a^+$ (including the truncation operation). We can use the Continuous Mapping Theorem [4] to claim that $\hat{\pi}_a^+ \xrightarrow{p} \pi_a \implies \hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.

(2) **Bias**: We will now proceed towards bounding the bias of the threshold for an arbitrary group $A = a$. We can break down our bias as follows: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a^+]| = |\mathbb{E}_{D_\alpha}[\alpha_a - \hat{\alpha}_a^+]| = |\mathbb{E}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}}] + \mathbb{E}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^C}]| \leq |\mathbb{E}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}}]| + |\mathbb{E}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^C}]|$, where $\mathcal{G}$ is defined next.

Let $\mathcal{G} = \{|\hat{\pi}_a - \pi_a| \leq \epsilon\}$ be the good event that our estimate of π_a is close. On this event, for sufficiently large N_α ($N_\alpha > \frac{2}{\pi_a}$), $\hat{\pi}_a^+ = \hat{\pi}_a$ and hence $\hat{\alpha}_a^+ = \hat{\alpha}_a$. Using Hoeffding's inequality, we know that $\hat{\pi}_a \in (\pi_a - \epsilon, \pi_a + \epsilon)$ with $1 - \delta/2$ probability over the choice of data sample D_α. We can set $\epsilon \leq \pi_a/2$, which means $N_\alpha > \frac{2}{\pi_a^2}\log(\frac{4}{\delta})$.

Consider the first term:

[4] stat.cmu.edu

$$
\begin{aligned}
|\mathbb{E}_{D_\alpha}\left[(\alpha_a - \hat{\alpha}_a)\mathbb{1}_{\mathcal{G}}\right]| &\leq \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a|\mathbb{1}_{\mathcal{G}}\right] = \mathbb{E}_{D_\alpha}\left[|\Pi(\alpha_a) - \Pi(\tilde{\alpha}_a)|\mathbb{1}_{\mathcal{G}}\right] \\
&\leq \mathbb{E}_{D_\alpha}\left[|\alpha_a - \tilde{\alpha}_a|\mathbb{1}_{\mathcal{G}}\right] \text{ (Since } \Pi \text{ is Lipschitz)} \\
&= \mathbb{E}_{D_\alpha}\left[\left|\frac{\lambda(2a-1)}{2\pi_a} - \frac{\lambda(2a-1)}{2\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&\leq \frac{|\lambda|}{2}\cdot\mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_a} - \frac{1}{\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&= \frac{|\lambda|}{2}\cdot\mathbb{E}_{D_\alpha}\left[\left|\frac{\hat{\pi}_a - \pi_a}{\pi_a\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] = \frac{|\lambda|}{2}\cdot\mathbb{E}_{D_\alpha}\left[\left|\frac{\hat{\pi}_a - \pi_a}{\pi_a\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&= \frac{|\lambda|}{2}\cdot\mathbb{E}_{D_\alpha}\left[\frac{|\hat{\pi}_a - \pi_a|}{\pi_a\hat{\pi}_a}\mathbb{1}_{\mathcal{G}}\right] \text{ (since } \pi_a, \hat{\pi}_a > 0) \\
&\leq \frac{|\lambda|}{\pi_a^2}\mathbb{E}_{D_\alpha}\left[|\hat{\pi}_a - \pi_a|\mathbb{1}_{\mathcal{G}}\right] \text{ (On } \mathcal{G}, \frac{1}{\pi_a\hat{\pi}_a} \leq \frac{2}{\pi_a^2}).
\end{aligned}
$$

Since $\hat{\pi}_a$ is an unbiased estimate of π_a, using Lemma B.2, $\mathbb{E}_{D_\alpha}\left[|\hat{\pi}_a - \pi_a|\right] \leq \sqrt{Var(\hat{\pi}_a)}$. Using $\mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a|\mathbb{1}_{\mathcal{G}}\right] \leq \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a|\right] \leq \frac{|\lambda|\sqrt{Var(\hat{\pi}_a)}}{\pi_a^2}$, we get:

$$
|\mathbb{E}_{D_\alpha}\left[(\alpha_a - \hat{\alpha}_a)\mathbb{1}_{\mathcal{G}}\right]| \leq \frac{|\lambda|\sqrt{Var(\hat{\pi}_a)}}{\pi_a^2} = \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}}
$$

where $Var(\hat{\pi}_a) = Var(\frac{1}{N_\alpha}\sum_{i\in[N_\alpha]}\mathbb{I}\left(a_i = a\right)) = \frac{1}{N_\alpha^2}Var(\sum_{i\in[N_\alpha]}\mathbb{I}\left(a_i = a\right)) = \frac{\pi_a(1-\pi_a)}{N_\alpha}$.

On $\mathcal{G}^c$, we can use the fact that both $\alpha_a, \hat{\alpha}_a^+ \in [0,1]$:

$$
\begin{aligned}
&|\mathbb{E}_{D_\alpha}\left[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^C}\right]| \leq |\mathbb{E}_{D_\alpha}\left[1\cdot\mathbb{1}_{\mathcal{G}^C}\right]| \\
&= \Pr\left(\mathcal{G}^C\right) \leq 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right) \text{ (Using Hoeffding's inequality on } \Pr\left(\mathcal{G}^{\mathcal{C}}\right)).
\end{aligned}
$$

Overall,

$$
|\alpha_a - \mathbb{E}_{D_\alpha}\left[\hat{\alpha}_a^+\right]| \leq \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right).
$$

(3) **Sample Complexity**: We now want to study the number of samples required for a good estimator of α_a^+. We want to ensure that $\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon\right) \leq \delta$. Note that $\{|\hat{\alpha}_a^+ - \alpha_a| > \epsilon\} \subseteq \{|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\}$.

Reusing the analysis from the previous part, we can first ensure that $\Pr\left(\mathcal{G}^C\right) = \Pr\left(|\pi_a - \hat{\pi}_a| > \pi_a/2\right) < \delta/2$ by using Hoeffding's inequality to obtain that $N_\alpha > \frac{2}{\pi_a^2}\log\frac{4}{\delta}$.

Next, as stated earlier, on the good event $\mathcal{G}$, as long as $N_\alpha > \frac{2}{\pi_a}$ (which means $1/N_\alpha < \pi_a/2$), $\hat{\pi}_a^+ = \hat{\pi}_a$ and hence $\hat{\alpha}_a^+ = \tilde{\alpha}_a$. This implies that $|\tilde{\alpha}_a^+ - \alpha_a| = |\frac{\lambda}{2}|\cdot|\frac{1}{\hat{\pi}_a} - \frac{1}{\pi_a}| \leq \frac{|\lambda|}{\pi_a^2}|\hat{\pi}_a - \pi_a|$.

Therefore, on $\mathcal{G}$, $\{|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\} \subseteq \{|\hat{\pi}_a - \pi_a| > \frac{\epsilon\pi_a^2}{|\lambda|}\}$. Now, we can study the entire event and upper bound its failure probability:

$$
\begin{aligned}
&\Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\right) \leq \Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon \cap \mathcal{G}\right) + \Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon \cap \mathcal{G}^C\right) \\
&\leq \Pr\left(|\hat{\pi}_a - \pi_a| > \frac{\epsilon\pi_a^2}{|\lambda|}\right) + \Pr\left(\mathcal{G}^C\right) \\
&\leq 2\exp(-\frac{2N_\alpha\epsilon^2\pi_a^4}{\lambda^2}) + 2\exp(-\frac{N_\alpha\pi_a^2}{2}) \text{ (Using Hoeffding's Inequality and earlier bound on } \mathcal{G}^C)
\end{aligned}
$$

Bounding the second term by $\delta/2$ gave us $N_\alpha > \frac{2}{\pi_a^2}\log\frac{4}{\delta}$. Doing the same for the first term gives us $N_\alpha > \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}$. Overall, $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$.

□

Theorem D.2. *(EOP results from Theorem 4.1 in the main text) (Consistency and Sample Complexity for EOP threshold) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and assuming $2 - \frac{(2a-1)\lambda}{\pi_{1a}} \geq \tau > 0$ (denominator bounded away from zero), for the case of equal opportunity, consider the following estimator of the threshold: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2-\frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}}$. Then the following holds:*

1. *Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.*
2. *Bias: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a^+]| \leq \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp(-2\epsilon^2 N_\alpha)$, where $\epsilon \leq \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$.*
3. *Sample Complexity: With $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr(|\hat{\alpha}_a - \alpha_a| > \epsilon) \leq \delta$.*

Proof. We will prove the results for an arbitrary $a \in \{0, 1\}$. Due to Lemma 3.8 and our projection operation Π on the empirical estimator, α_a and $\hat{\alpha}_a^+$ are both restricted to the interval $[0, 1]$. Using this, note that in our domain of $\lambda \in [0, \lambda_{\max}]$ or $\lambda \in [-\lambda_{\min}, 0]$, $\alpha_a = \Pi(\alpha_a)$, which is a useful property we exploit throughout the proof.

(1) **Consistency**: By the weak law of large numbers, $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$. Consider $\hat{\pi}_{1a}^+$. Since $\pi_{1a} > 0$ by assumption, we can pick a sufficiently large N_α such that $1/N_\alpha \leq \pi_{1a}/2$. Therefore the event $\{\hat{\pi}_{1a} < 1/N_\alpha\}$ is contained in the event $\{\hat{\pi}_{1a} < \pi_{1a}/2\}$ which is further contained in $\{|\hat{\pi}_{1a} - \pi_{1a}| \geq \pi_{1a}/2\}$. Since $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$, $\Pr(|\hat{\pi}_{1a} - \pi_{1a}| \geq \pi_{1a}/2) \to 0$ and hence $\Pr(\hat{\pi}_{1a}^+ \neq \hat{\pi}_{1a}) \to 0$, and hence $\hat{\pi}_{1a}^+ \xrightarrow{p} \pi_{1a}$.

Now, we will show consistency in two steps. First, let the denominator be defined as $\tilde{d}_a = 2 - \frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}$and $d_a = 2 - \frac{(2a-1)\lambda}{\pi_{1a}}$. Since $\pi_{1a} > 0$ by assumption, and $\tilde{\pi}_{1a}^+ \xrightarrow{p} \pi_{1a}$, we can use the Continuous Mapping Theorem to claim that $\tilde{d}_a \xrightarrow{p} d_a$.

Again, since $d_a \geq \tau > 0$, then by application of the Continuous Mapping Theorem, $\frac{1}{\tilde{d}_a} \xrightarrow{p} \frac{1}{d_a}$ and hence $\hat{\alpha}_a \xrightarrow{p} \alpha_a$ (as the truncation operation Π is also continuous).

(2) Bias: We want to now bound the bias of the estimator: $|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a]|$:

$$\begin{aligned}
|\alpha_a - \mathbb{E}_{D_\alpha}[\hat{\alpha}_a]| &= |\mathbb{E}_{D_\alpha}[\alpha_a - \hat{\alpha}_a]| \leq \mathbb{E}_{D_\alpha}[|\alpha_a - \hat{\alpha}_a|] = \mathbb{E}_{D_\alpha}[|\Pi(\alpha_a) - \Pi(\tilde{\alpha}_a)|] \\
&\leq \mathbb{E}_{D_\alpha}[|\alpha_a - \tilde{\alpha}_a|] \quad \text{(Since } \Pi \text{ is Lipschitz)} \\
&= \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right|\right] = \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right| \mathbb{1}_{\mathcal{G}}\right] + \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right| \mathbb{1}_{\mathcal{G}^C}\right],
\end{aligned}$$

where we let $\mathcal{G} = \{|\hat{\pi}_{1a} - \pi_{1a}| \leq \epsilon\}$ denote the good event that $\hat{\pi}_{1a}$ is close to π_{1a}. On this event, for sufficiently large N_α ($N_\alpha > \frac{2}{\pi_{1a}}$), $\hat{\pi}_{1a}^+ = \hat{\pi}_{1a}$ and hence $\hat{\alpha}_a^+ = \hat{\alpha}_a$. We can set $\epsilon \leq \pi_{1a}/2$ so that $\hat{\pi}_{1a} \geq \pi_{1a}/2$.

On $\mathcal{G}$:

$$|d_a - \tilde{d}_a| \leq |\lambda|\left|\frac{\hat{\pi}_{1a} - \pi_{1a}}{\pi_{1a}\hat{\pi}_{1a}}\right| \leq \frac{2|\lambda|}{\pi_{1a}^2}|\hat{\pi}_{1a} - \pi_{1a}|$$

Similarly, we can also bound $\tilde{d}_a$ on $\mathcal{G}$:

$$\tilde{d}_a \geq d_a - |\tilde{d}_a - d_a| \geq \tau - \frac{2|\lambda|\epsilon}{\pi_{1a}^2}.$$

We can ensure that $\tilde{d}_a \geq \tau/2$ on $\mathcal{G}$, which means $\epsilon \leq \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$. Finally, for the first term of the bias bound, on the good event $\mathcal{G}$:

$$\mathbb{E}_{D_\alpha}\left[\Big|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\Big|\mathbb{1}_{\mathcal{G}}\right] = \mathbb{E}_{D_\alpha}\left[\frac{|\tilde{d}_a - d_a|}{d_a\tilde{d}_a}\mathbb{1}_{\mathcal{G}}\right] | \leq \frac{2}{\tau^2}\cdot\frac{2|\lambda|}{\pi_{1a}^2}\mathbb{E}_{D_\alpha}\left[|\hat{\pi}_{1a} - \pi_{1a}|\right].$$

On $\mathcal{G}^c$, we can use the fact that both $\alpha_a, \hat{\alpha}_a^+ \in [0,1]$:

$$\begin{aligned}&\mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{1}_{\mathcal{G}^C}\right] \leq \mathbb{E}_{D_\alpha}\left[1\cdot\mathbb{1}_{\mathcal{G}^C}\right]\\ &= \Pr\left(\mathcal{G}^C\right) \leq 2\exp(-2\epsilon^2 N_\alpha) \text{ (Using Hoeffding's inequality on } \Pr\left(\mathcal{G}^{\mathcal{C}}\right)).\end{aligned}$$

Overall,

$$|\alpha_a - \mathbb{E}_{D_\alpha}\left[\hat{\alpha}_a^+\right]| \leq \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\mathbb{E}_{D_\alpha}\left[|\hat{\pi}_{1a} - \pi_{1a}|\right] + 2\exp(-2\epsilon^2 N_\alpha), \text{ where } \epsilon \leq \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\right\}.$$

To further simplify this, using Lemma B.2, we know that $\mathbb{E}\left[|\pi_{1a} - \hat{\pi}_{1a}|\right] \leq \sqrt{Var(\hat{\pi}_{1a})} = \sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}}$, and hence:

$$|\alpha_a - \mathbb{E}_{D_\alpha}\left[\hat{\alpha}_a^+\right]| \leq \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp(-2\epsilon^2 N_\alpha), \text{ where } \epsilon \leq \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\right\}.$$

(3) **Sample Complexity**: For sample complexity, we can proceed by returning a high probability bound on the good event derived in the previous part. Note that $\{|\hat{\alpha}_a^+ - \alpha_a| > \epsilon\} \subseteq \{|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\}$.

Recall that $\mathcal{G} = \{|\hat{\pi}_{1a} - \pi_{1a}| \leq \epsilon\}$ denotes the good event that $\hat{\pi}_{1a}$ is close to π_{1a}. We can set $\epsilon \leq \pi_{1a}/2$ so that $\hat{\pi}_{1a} \geq \pi_{1a}/2$, whenever $1/N_\alpha \leq \pi_{1a}/2 \implies N_\alpha \geq 2/\pi_{1a}$.

From the previous part, we know that on $\mathcal{G}$:

- $|d_a - \tilde{d}_a| \leq \frac{2|\lambda|\epsilon}{\pi_{1a}^2}$,
- $\tilde{d}_a \geq \tau/2$ for $\epsilon \leq \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$, and
- Using both of the results above, $|\alpha_a - \tilde{\alpha}_a| \leq \frac{4|\lambda|\epsilon}{\tau^2\pi_{1a}^2}$ on $\mathcal{G}$.

Finally, if we want $|\tilde{\alpha}_a - \alpha_a| \leq \epsilon'$, then $\frac{4|\lambda|\epsilon}{\tau^2\pi_{1a}^2} \leq \epsilon' \implies \epsilon \leq \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}$. Combining this with the previous min bound on ϵ, we get $\epsilon = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.

We have now shown that $\{|\hat{\pi}_{1a} - \pi_{1a}| \leq \epsilon\} \subseteq \{|\alpha_a - \hat{\alpha}_a| \leq \epsilon'\}$. Taking complements and probabilities on both sides, we get:

$$\Pr\left(|\alpha_a - \hat{\alpha}_a| > \epsilon'\right) \leq \Pr\left(|\hat{\pi}_{1a} - \pi_{1a}| > \epsilon\right) \leq 2\exp(-2N_\alpha\epsilon^2) \leq \delta,$$

which gives us $N_\alpha \geq \frac{1}{2\epsilon^2}\log(\frac{2}{\delta})$ for $\epsilon = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.

Refactoring notations (using t for ϵ and ϵ for ϵ'), with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leq \delta$.

□

Theorem D.3. *(Theorem 4.3 in the main text) Consider the Bayes optimal classifier for the cost-sensitive risk with group-dependent costs α_a:*

$$g^*_\alpha(x,a) = \begin{cases} 1, & \text{if } \eta(x,a) \geq \alpha_a \\ 0, & \text{otherwise.} \end{cases}$$

1. *Bayes Risk: The Bayes error with respect to the cost-sensitive risk can be expressed as $\beta_{CS} = \mathbb{E}_{X,A}\left[\min((1-\alpha_A)\eta(X,A), \alpha_A(1-\eta(X,A)))\right]$.*
2. *Consistency: $\hat{\beta}_{CS} = \frac{1}{N_\beta}\sum_{i\in[N_\beta]}\min((1-\alpha_i)c_i, \alpha_i(1-c_i))$ is a consistent estimator of β_{CS}, where for the i^{th} instance (c_i, a_i), $c_i = \eta(x_i, a_i)$ and $\alpha_i = \alpha_{a_i}$, the group dependent cost.*
3. *Sample Complexity: For any $\delta > 0$, with probability at least $1-\delta$, $|\hat{\beta}_{CS} - \beta_{CS}| \leq \frac{1}{4}\sqrt{\frac{1}{N_\beta}\log(\frac{2}{\delta})}$.*

Proof. (1) **Bayes Risk**: We will adapt the standard proof technique[5] of the Bayes error for cost-sensitive settings. The expected α-cost sensitive risk for the classifier g^* can be expressed as:

$$\begin{aligned}
R_\alpha(g^*) &= \sum_{a\in\mathcal{A}}\left((1-\alpha_a)\cdot\mathbb{E}_{X,Y=1,A=a}\left[l_{01}(g^*(X,a),1)\right] + \alpha_a\cdot\mathbb{E}_{X,Y=0,A=a}\left[l_{01}(g^*(X,a),0)\right]\right) \\
&= \sum_{a\in\mathcal{A}}\mathbb{E}_{X,a}\left[\mathbb{E}_{Y=1|X,a}\left[(1-\alpha_a)\mathbb{I}\left(Y=1, g^*(X,a)=0\right)\right] + \mathbb{E}_{Y=0|X,a}\left[\alpha_a\mathbb{I}\left(Y=0, g^*(X,a)=1\right)\right]\right] \\
&= \sum_{a\in\mathcal{A}}\mathbb{E}_{X,a}\left[(1-\alpha_a)\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\eta(X,a) + \alpha_a\mathbb{I}\left(\eta\geq\alpha_a\right)(1-\eta(X,a))\right]
\end{aligned}$$

At this point, note that $\eta < \alpha_a \implies (1-\alpha_a)\eta < \alpha_a(1-\eta)$ and similarly $\eta \geq \alpha_a \implies \alpha_a(1-\eta) \leq (1-\alpha_a)\eta$. Therefore,

$$\begin{aligned}
R_\alpha(g^*) &= \sum_{a\in\mathcal{A}}\mathbb{E}_{X,a}\left[(1-\alpha_a)\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\eta(X,a) + \alpha_a\mathbb{I}\left(\eta\geq\alpha_a\right)(1-\eta(X,a))\right] \\
&= \sum_{a\in\mathcal{A}}\mathrm{E}_{X,a}[\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\min((1-\alpha_a)\eta(X,a), \alpha_a(1-\eta(X,a))) \\
&\qquad + \mathbb{I}\left(\eta(X,a)\geq\alpha_a\right)\min(\alpha_a(1-\eta(X,a)), \eta(X,a)(1-\alpha_a))] \\
&= \mathbb{E}_{X,A}\left[\min((1-\alpha_A)\eta(X,A), \alpha_A(1-\eta(X,A)))\right].
\end{aligned}$$

It remains to show that g^* is the Bayes optimal classifier. Let g' be any arbitrary classifier. Then,

[5] ocw.mit.edu

$$
\begin{aligned}
R_\alpha(g') - R_\alpha(g^*) &= \mathbb{E}_{X,a}\left[(1-\alpha_a)\eta(X,a)\mathbb{I}\left(g'(X,a)=0\right) + \alpha_a(1-\eta(X,a))\mathbb{I}\left(g'(X,a)=1\right)\right] \\
&\quad - \mathbb{E}_{X,a}\left[(1-\alpha_a)\eta(X,a)\mathbb{I}\left(g^*(X,a)=0\right) + \alpha_a(1-\eta(X,a))\mathbb{I}\left(g^*(X,a)=1\right)\right] \\
&= \mathrm{E}_{X,a}[(\mathbb{I}\left(g'=0\right) - \mathbb{I}\left(g^*=0\right))(1-\alpha_a)\eta(X,a) \\
&\qquad\qquad + (\mathbb{I}\left(g'=1\right) - \mathbb{I}\left(g^*=1\right))\alpha_a(1-\eta(X,a))] \\
&= \mathbb{E}_{X,a}\left[(\mathbb{I}\left(g'=0\right) - \mathbb{I}\left(g^*=0\right))(\eta(X,a)-\alpha_a)\right] \\
&= \mathbb{E}_{X,a}\left[\mathbb{I}\left(g' \neq g^*\right)\cdot(\eta(X,a)-\alpha_a)\cdot sgn(\eta(X,a)-\alpha_a)\right], \\
&\qquad\qquad \text{using the fact that } g^* = \mathbb{I}\left(\eta \geq \alpha_a\right) \\
&= \mathbb{E}_{X,a}\left[\mathbb{I}\left(g' \neq g^*\right)\cdot|\eta(X,a)-\alpha_a|\right] \geq 0.
\end{aligned}
$$

(2) **Consistency**:

$$
\begin{aligned}
\mathbb{E}_{(c_i,a_i)_{i=1}^{N_\beta}}\left[\hat{\beta}_{CS}\right] &= \mathbb{E}_{(c_i,a_i)_{i=1}^{N_\beta}}\left[\frac{1}{N_\beta}\sum_{i\in[N_\beta]}\min((1-\alpha_i)c_i, \alpha_i(1-c_i))\right] \\
&= \frac{1}{N_\beta}\sum_{i\in[N_\beta]}\mathbb{E}_{(c_i,a_i)\sim D}\left[\min((1-\alpha_i)c_i, \alpha_i(1-c_i))\right] \\
&= \beta_{CS}.
\end{aligned}
$$

By the weak law of large numbers, $\hat{\beta}_{CS} \xrightarrow{p} \beta_{CS}$.

(3) **Sample Complexity**: The proof for sample complexity is a straightforward application of Hoeffding's inequality, by noting that $\min((1-\alpha_i)c_i, \alpha_i(1-c_i)) \in [0, 1/4]$:

$$
\begin{aligned}
&\Pr\left(|\hat{\beta}_{CS} - \beta_{CS}| \geq \epsilon\right) \leq 2\exp\left(\frac{-2N_\beta\epsilon^2}{(1/4)^2}\right) \leq \delta \\
&\implies \epsilon \geq \frac{1}{4}\sqrt{\frac{1}{2N_\beta}\log(\frac{2}{\delta})}.
\end{aligned}
$$

□

Theorem D.4. *(Theorem 4.4 in the main text) Suppose the dataset D_α is used to construct a consistent estimator of the cost-sensitive threshold α_a (Theorem 4.1). Then, given a dataset D_β the following statements are true:*

1. *Consistency: $\hat{\beta}_{CS,\hat{\alpha}} = \frac{1}{N_\beta}\sum_{i\in[D_\beta]}\min((1-\hat{\alpha}^+_{a_i})c_i, \hat{\alpha}^+_{a_i}(1-c_i))$ is a consistent estimator of β_{CS}.*
2. *Bias: For a fixed $\delta \in (0,1)$:*
   - *Demographic Parity: $|\beta_{CS} - \mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(|\lambda|\cdot\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,*
   - *Equal Opportunity: $|\beta_{CS} - \mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(\frac{|\lambda|}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.*
3. *Sample Complexity: With $1-\delta$ probability, $|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| < \epsilon$ as long as $N_\beta \geq \frac{1}{8\epsilon^2}\log(\frac{4}{\delta})$ and:*
   - *Demographic Parity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,*
   - *Equal Opportunity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\tau^4\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.*

Proof. (1) **Consistency**: For showing consistency, note that conditioned on D_α, our thresholds $\hat{\alpha}_a^+$ are fixed, and hence by Theorem 4.3, $\hat{\beta}_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS,\hat{\alpha}}$, where $\beta_{CS,\hat{\alpha}}$ is the cost sensitive Bayes risk with respect to group-aware thresholds $\hat{\alpha}_a^+, \forall a \in \mathcal{A}$. It remains to show that $\beta_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

From Theorem 4.1, $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$. Note that $f(x) = \mathbb{E}\left[min\{(1-x)\eta, x(1-\eta)\}\right]$ is a continuous function of x. Hence, by the Continuous Mapping Theorem, $\beta_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

(2) **Bias**:

$$
\begin{aligned}
|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| &= |\mathbb{E}_{D_\beta, D_\alpha}\left[\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}\right]| \\
&= |\mathbb{E}_{D_\alpha}\left[\mathbb{E}_{D_\beta}\left[(\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}})|D_\alpha\right]\right]| = |\mathbb{E}_{D_\alpha}\left[\beta_{CS} - \beta_{CS,\hat{\alpha}}\right]| \\
&\leq \mathbb{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}}|\right].
\end{aligned}
$$

At this point, consider the following:

$$
\begin{aligned}
|\beta_{CS,x} - \beta_{CS,y}| &= |\mathbb{E}_{X,A}\left[\min(\eta(1-x), x(1-\eta)) - \min(\eta(1-y), y(1-\eta))\right]| \\
&= |\mathbb{E}\left[\frac{\eta+x}{2} - x\eta - |\frac{x-\eta}{2}| - \frac{\eta+y}{2} + y\eta + |\frac{y-\eta}{2}|\right]| \\
&\text{(Using } \min(a,b) = \frac{a+b}{2} - |\frac{a-b}{2}|) \\
&= \frac{1}{2}|\mathbb{E}\left[(x-y)(1-2\eta) - |\eta - x| + |\eta - y|\right]| \\
&= \frac{1}{2}|(x-y)(1-2\mathbb{E}\left[\eta\right]) + \mathbb{E}\left[|\eta - y| - |\eta - x|\right]| \\
&\leq \frac{1}{2}|x-y| \cdot |1 - 2\mathbb{E}\left[\eta\right]| + \frac{1}{2}\mathbb{E}\left[||\eta - y| - |\eta - x||\right] \\
&\leq \frac{1}{2}|x-y| \cdot |1 - 2\mathbb{E}\left[\eta\right]| + \frac{1}{2}\mathbb{E}\left[|x-y|\right] \text{(Using Reverse Triangle Inequality)} \\
&= \frac{1}{2}|1 - 2\mathbb{E}\left[\eta\right]||x-y| + \frac{1}{2}|x-y| = \frac{|x-y|}{2}(1 + |1-2\pi|) \\
&\leq |x-y| \text{(Since } \pi \in [0,1]).
\end{aligned}
$$

The above result shows that the Bayes risk is Lipschitz in the threshold argument α with $L = 1$. This result will be useful later.

We know from Theorem 4.1 that the bias of $\hat{\alpha}_a^+$ is bounded: $|\mathbb{E}\left[\hat{\alpha}_a^+\right] - \alpha_a| \leq U_m$, where U_m depends on the fairness metric. Therefore, setting $x = \mathbb{E}_{D_\alpha}\left[\hat{\alpha}\right]$ and $y = \alpha_a$ yields us:

$$
\begin{aligned}
|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| &\leq \mathbb{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}}|\right] \\
&\leq \pi_0 \mathbb{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}_0^+}|\right] + \pi_1 \mathbb{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}_1^+}|\right] \\
&\leq \pi_0 \mathbb{E}_{D_\alpha}\left[|\alpha_0 - \hat{\alpha}_0^+|\right] + \pi_1 \mathbb{E}_{D_\alpha}\left[|\alpha_1 - \hat{\alpha}_1^+|\right]
\end{aligned}
$$

From Theorem 4.1, we know that with $1 - \delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$. Using this, we get:

$$
\begin{aligned}
&|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| \leq \pi_0 \mathbb{E}_{D_\alpha}\left[|\alpha_0 - \hat{\alpha}_0^+|\right] + \pi_1 \mathbb{E}_{D_\alpha}\left[|\alpha_1 - \hat{\alpha}_1^+|\right] \\
&= \sum_{a\in\{0,1\}} \pi_a \left(\mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right)\right] + \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon\right)\right]\right) \\
&\leq \sum_{a\in\{0,1\}} \pi_a \left(\epsilon \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon\right)\right) \quad \text{(Since } |\alpha_a - \hat{\alpha}_a^+| \leq 1\text{)} \\
&\leq \sum_{a\in\{0,1\}} \pi_a(\epsilon_a(N_\alpha, \delta_a) + \delta_a).
\end{aligned}
$$

where, $\epsilon_a(N_\alpha, \delta_a)$ can de derived from the respective sample complexities in Theorem 4.1.

For DP, using the dominant term, $\epsilon_a = \frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$ and setting $\delta = \max\{\delta_0, \delta_1\}$, giving us:

$$
|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(|\lambda| \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).
$$

For EOP, using the ϵ dependent term, $\epsilon_a = \frac{|\lambda|}{\tau^2 \pi_{1a}^2}\sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}}$ and setting $\delta = \max\{\delta_0, \delta_1\}$, giving us:

$$
|\beta_{CS} - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(\frac{|\lambda|}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).
$$

(3) **Sample Complexity**: We want to ensure the following:

$$
\Pr\left(|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| > \epsilon\right) \leq \delta.
$$

To do so, consider the following breakdown: $|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| \leq |\beta_{CS} - \beta_{CS,\hat{\alpha}}| + |\beta_{CS,\hat{\alpha}} - \hat{\beta}_{CS,\hat{\alpha}}|$.

Consider the random error: $|\beta_{CS,\hat{\alpha}} - \hat{\beta}_{CS,\hat{\alpha}}|$. This will give us a sample complexity lower bound for N_β. Bounding this is a simple application of Hoeffding's inequality with $X_i = \min\{(1 - \hat{\alpha}_{a_i}^+)c_i, \hat{\alpha}_{a_i}^+(1 - c_i)\} \in [0, 1/4]$:

$$
\Pr\left(|\beta_{CS,\hat{\alpha}_a} - \hat{\beta}_{CS,\hat{\alpha}}| > \epsilon/2\right) \leq 2\exp(-8N_\beta\epsilon^2) \leq \delta/2 \implies N_\beta \geq \frac{1}{8\epsilon^2}\log(\frac{4}{\delta})
$$

Now consider the error due to threshold estimation. We know from our analysis of bias that $|\beta_{CS,x} - \beta_{CS,y}| \leq |x - y| \implies |\beta_{CS} - \beta_{CS,\hat{\alpha}}| \leq \sum_{a\in\{0,1\}} \pi_a \cdot |\hat{\alpha}_a^+ - \alpha_a|$.

From the sample complexity results in Theorem 4.1, we can derive the following: with $\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/2\right) \leq U_m(\epsilon/2) \leq \delta/2$, where U_m depends on the fairness metric we choose. This will provide us with a bound on $N_\alpha \geq N_\alpha(m)$.

For demographic parity, again using the dominant term:

$$
\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/4\pi_a\right) \leq \delta/4 \implies N_\alpha \geq \frac{8\lambda^2}{\epsilon^2\pi_a^2}\log\left(\frac{16}{\delta}\right) = \mathcal{O}\left(\frac{\lambda^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right),
$$

For equal opportunity, recall that with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leq \delta$. Using these bounds with our new ϵ and δ we get:

$$\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/4\pi_a\right) \leq \delta/4 \implies N_\alpha \geq \frac{1}{2}\log\left(\frac{8}{\delta}\right)\max\left\{\frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau\pi_{1a}^2}, \frac{4|\lambda|}{\epsilon\tau^2\pi_{1a}^2}\right\}^2$$
$$= \mathcal{O}\left(\frac{\lambda^2}{\tau^4\epsilon^2}\log\left(\frac{1}{\delta}\right)\right).$$

□

Proof for Theorem 4.5: Recall from Definition 3.5 the definition of expected risk:

$$R(g) = \sum_{a\in\{0,1\}} \pi_a \cdot \left(\mathbb{E}_{X|A=a}\left[(1-2\eta(X,a))g(X,a)\right]\right) + \pi$$
$$= \mathbb{E}_{X,A}\left[(1-2C)\mathbb{I}\left(C \geq \alpha_A\right)\right] + \pi$$

A finite sample estimator of the above risk can be written down as follows:

$$\hat{R}(\hat{g}) = \frac{1}{N_\beta}\sum_{i\in[D_\beta]}(1-2c_i)\mathbb{I}\left(c_i \geq \hat{\alpha}_{a_i}^+\right) + \hat{\pi}$$

Theorem D.5. *(Proof for Theorem 4.5 from the main text) Consider $\hat{R}(\hat{g})$, the empirical estimator for the expected risk from Definition 3.5. Then, the following properties hold:*

1. *Consistency: $\hat{R}(\hat{g}) \xrightarrow{p} R(g)$.*
2. *Bias: For a fixed $\delta \in (0,1)$:*
   - *Demographic Parity: $|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(|\lambda| \cdot \max\{L_0, L_1\} \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,*
   - *Equal Opportunity: $|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda|\max\{L_0,L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.*
3. *Sample Complexity: With $1-\delta$ probability, $|R(g) - \hat{R}(\hat{g})| < \epsilon$ as long as $N_\beta \geq \frac{18}{\epsilon^2}\log\left(\frac{6}{\delta}\right)$ and:*
   - *Demographic Parity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,*
   - *Equal Opportunity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$.*

Proof. To prove the above results, we will use an intermediate estimator $R(\hat{g}) = \mathbb{E}_{X,A}\left[(1-2C)\mathbb{I}\left(C \geq \hat{\alpha}_A^+\right)\right] + \pi$.

1. **Consistency**: We want to show the following result:

$$\Pr\left(|R(g) - \hat{R}(\hat{g})| > \epsilon\right) \xrightarrow{p} 0 \text{ as } N_\alpha, N_\beta \to \infty.$$

We can decompose the consistency gap as follows:

$$|R(g) - \hat{R}(\hat{g})| \leq |R(g) - R(\hat{g})| + |R(\hat{g}) - \hat{R}(\hat{g})|$$

Since D_α and D_β are independently drawn samples, conditioned on D_α, $\hat{\alpha}_a^+$ is a constant for all $A = a$ and hence $\hat{R}(\hat{g}) \xrightarrow{p} R(\hat{g})$ by the weak law of large numbers. It remains to show how $R(\hat{g}) \xrightarrow{p} R(g)$.

Let $\phi_a(t) = \mathbb{E}_{X|A=a}\left[(1-2C)\mathbb{I}\left(C \geq t\right)|A = a\right], t \in (0,1)$, so that $R(g) = \sum_{a\in\{0,1\}} \pi_a \cdot \phi_a(\alpha_a)$. Let $0 \leq u < v \leq 1$ and consider the following analysis for a fixed $A = a$:

$$\begin{aligned}
|\phi_a(u) - \phi_a(v)| &= |\mathbb{E}\left[(1-2C)(\mathbb{I}\left(C \geq u\right) - \mathbb{I}\left(C \geq v\right))|A = a\right]| \\
&\leq \Pr\left(u \leq C < v | A = a\right) \text{ (Since } (\mathbb{I}\left(C \geq u\right) - \mathbb{I}\left(C \geq v\right)) \neq 0 \text{ when } C \in [u, v)) \\
&= |F_a(v) - F_a(u)|
\end{aligned}$$

Now, from Theorem 4.1, we know that $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ and since F_a is continuous due to Assumption 2.1 and $\alpha_a > 0$, by the continuous mapping theorem, $\phi(\hat{\alpha}_a^+) \xrightarrow{p} \phi(\alpha_a)$. Since $R(g) = \sum_{a\in\{0,1\}} \pi_a \cdot \phi_a(\alpha_a)$, $R(\hat{g}) \xrightarrow{p} R(g)$.

2. **Bias**: We want to bound the following quantity: $|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]|$. Let $R_0(g) = R(g) - \pi$, and similarly define $R_0(\hat{g}) = R(\hat{g}) - \pi$.

Note that since D_α and D_β are independently sampled, $\mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right] = \mathbb{E}_{D_\alpha}\left[R_0(\hat{g})\right] + \mathbb{E}_{D_\alpha}\left[\hat{\pi}\right]$, which means:

$$\begin{aligned}
|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| &\leq |R_0(g) - \mathbb{E}_{D_\alpha}\left[R_0(\hat{g})\right]| + |\pi - \mathbb{E}_{D_\alpha}\left[\hat{\pi}\right]| \\
&= |R_0(g) - \mathbb{E}_{D_\alpha}\left[R_0(\hat{g})\right]| \text{ (Since } \hat{\pi} \text{ is an unbiased estimate of } \pi)
\end{aligned}$$

Let $\phi_a(t) = \mathbb{E}_{X|A=a}\left[(1-2C)\mathbb{I}\left(C \geq t\right)|A = a\right]$. From our consistency proof, we know that $|\phi_a(u) - \phi_a(v)| \leq |F_a(v) - F_a(u)|$. Using this:

$$\begin{aligned}
&|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| \leq |R_0(g) - \mathbb{E}_{D_\alpha}\left[R_0(\hat{g})\right]| = |\sum_{a\in\{0,1\}} \pi_a \mathbb{E}_{D_\alpha}\left[\phi_a(\alpha_a) - \phi_a(\hat{\alpha}_a^+)\right]| \\
&\leq \sum_{a\in\{0,1\}} \pi_a \mathbb{E}_{D_\alpha}\left[|\phi_a(\alpha_a) - \phi_a(\hat{\alpha}_a^+)|\right] \leq \sum_{a\in\{0,1\}} \pi_a L_a \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] \text{ (Using Assumption 2.2)}
\end{aligned}$$

From Theorem 4.1, we know that with $1-\delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$ We also know that $|\alpha_a - \hat{\alpha}_a^+| \leq 1$. Using this, we get:

$$\begin{aligned}
&\mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] = \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right)\right] + \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon\right)\right] \\
&\leq \epsilon \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon\right) \text{ (Since } |\alpha_a - \hat{\alpha}_a^+| \leq 1) \leq \epsilon + \delta.
\end{aligned}$$

and therefore,

$$|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| \leq \sum_{a\in\{0,1\}} \pi_a L_a \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] \leq \sum_{a\in\{0,1\}} \pi_a L_a (\epsilon_a + \delta_a).$$

For DP, using the sample complexity bound from Theorem 4.1 and the dominant term, $\epsilon_a = \frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$, and setting $\delta = \min\{\delta_0, \delta_1\}$:

$$|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(|\lambda| \cdot \max\{L_0, L_1\} \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).$$

For EOP, using the sample complexity bound from Theorem 4.1 and the ϵ dependent term, $\epsilon_a = \frac{|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{8\log(4/\delta_a)}{N_\alpha}}$, and setting $\delta = \min\{\delta_0, \delta_1\}$:

$$|R(g) - \mathbb{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda| \max\{L_0, L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).$$

(3) **Sample Complexity**: We want to ensure the following:

$$\Pr\left(|R(g) - \hat{R}(\hat{g})| > \epsilon\right) \leq \delta.$$

Note that we can decompose the absolute error as follows via the triangle inequality:

$$|R(g) - \hat{R}(\hat{g})| \leq |R_0(g) - R_0(\hat{g})| + |R_0(\hat{g}) - \hat{R}_0(\hat{g})| + |\hat{\pi} - \pi|.$$

Applying the union bound, we get:

$$\begin{aligned}\Pr\left(|R(g) - \hat{R}(\hat{g})| > \epsilon\right) \leq \Pr\left(|R_0(g) - R_0(\hat{g})| > \epsilon/3\right) + \Pr\left(|R_0(\hat{g}) - \hat{R}_0(\hat{g})| > \epsilon/3\right) \\ + \Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right).\end{aligned}$$

Conditioned on D_α, $\hat{\alpha}_a^+$ is a constant and hence we can control $\Pr\left(|R_0(\hat{g}) - \hat{R}_0(\hat{g})| > \epsilon/3\right)$ via the Hoeffding's inequality. Note that the random variable $Z_i = (1 - 2c_i)\mathbb{I}\left(c_i \geq \hat{\alpha}_{a_i}^+\right) \in [-1, 1]$.

$$\Pr\left(|R_0(\hat{g}) - \hat{R}_0(\hat{g})| > \epsilon/3 | D_\alpha\right) \leq \delta/3 \implies N_\beta \geq \frac{18}{\epsilon^2}\log\left(\frac{6}{\delta}\right).$$

Similarly, we can also control $\Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right)$:

$$\Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right) \leq \delta/3 \implies N_\alpha \geq \frac{9}{2\epsilon^2}\log\left(\frac{6}{\delta}\right)$$

Finally, to control $\Pr\left(|R_0(g) - R_0(\hat{g})| > \epsilon/3\right)$, we can borrow our analysis from the previous proof of bias:

$$|R_0(g) - R_0(\hat{g})| \leq \sum_{a \in \{0,1\}} \pi_a L_a |\alpha_a - \hat{\alpha}_a^+|.$$

By union bound and splitting the error budget between each group-dependent term (since there are two groups, if the sum exceeds $\epsilon/3$, at least one term exceeds $\epsilon/6$):

$$\Pr\left(|R_0(g) - R_0(\hat{g})| > \epsilon/3\right) \leq \sum_{a\in\{0,1\}} \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right).$$

For demographic parity, again using the dominant term:

$$\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right) \leq \delta/6 \implies N_\alpha \geq \frac{18\lambda^2 L_a^2}{\epsilon^2\pi_a^2}\log\left(\frac{24}{\delta}\right) = \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right).$$

For equal opportunity, recall that with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leq \delta$. Using these bounds with our new ϵ and δ we get:

$$\begin{aligned}\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right) \leq \delta/6 \implies N_\alpha &\geq \frac{1}{2}\log\left(\frac{12}{\delta}\right)\max\left\{\frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau\pi_{1a}^2}, \frac{24|\lambda|\pi_a L_a}{\epsilon\tau^2\pi_{1a}^2}\right\}^2 \\ &= \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

Taking a max over L_a completes our proof. □

Again, for convenience and readability, we will prove the DP and EOP results for Theorem 4.6.

Theorem D.6. *(Proof of the DP part of Theorem 4.6 in the main text) Suppose a dataset D_β is used to construct $\hat{\Delta}_{DP}(g_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the DP threshold (Theorem 4.1). Then the following holds:*

1. *Consistency: $\hat{\Delta}_{DP}(g_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{DP}(g_\alpha)$).*
2. *Bias: For $\delta \in (0,1)$, $|\Delta_{DP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]| = \mathcal{O}\left(\frac{1}{\sqrt{N_\beta}} + |\lambda|\cdot\max\{L_0, L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right)$.*
3. *Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{g}) - \Delta(g)| \leq \epsilon$, as long as $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log(\frac{1}{\delta})\right)$ and $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\cdot\max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.*

Proof. For the sake of this proof, we will denote $\Delta_{DP}(g_\alpha)$ as $\Delta(g)$ and $\hat{\Delta}_{DP}(g_{\hat{\alpha}_a^+})$ as $\hat{\Delta}(\hat{g})$. Then an intermediate estimator $\hat{\Delta}_{DP}(g_{\alpha_a})$ can be denoted as $\hat{\Delta}(g)$. We similarly overload the notations in this fashion for the positive rate estimator $\rho_{\mathrm{PR},a}$ with ρ_a.

(1) **Consistency**: Most of the inequality arguments in the proof are a direct consequence of the triangle inequality, by introducing an intermediate estimate of the PR ($\hat{\rho}_a(g)$).

We want to show that $\Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(g)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$. By Slutsky's theorem, it will be enough to upper-bound our target event as a sum of simpler events and show that if each of them converges in probability to zero, then we can claim our final result.

Consider the absolute difference term and apply the reverse triangle inequality ($||A|-|B|| \leq |A-B|$):

$$\begin{aligned}&|\Delta(g) - \hat{\Delta}(\hat{g})| \leq |\rho_1(g) - \rho_0(g) - \hat{\rho}_1(\hat{g}) + \hat{\rho}_0(\hat{g})| \\ &\leq |\rho_1(g) - \hat{\rho}_1(g)| + |\rho_0(g) - \hat{\rho}_0(g)| + |\hat{\rho}_1(g) - \hat{\rho}_1(\hat{g})| + |\hat{\rho}_0(g) - \hat{\rho}_0(\hat{g})|.\end{aligned}$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$.

Consider $|\rho_a(g) - \hat{\rho}_a(g)|$. By the Weak law of large numbers, $\frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} \mathbb{I}(c_i \geq \alpha_a) \xrightarrow{p} \mathbb{E}_{X|A=a}[\mathbb{I}(c_a \geq \alpha_a)]$, and hence $|\rho_a(g) - \hat{\rho}_a(g)| \xrightarrow{p} 0$.

Next, consider the term $|\hat{\rho}_a(g) - \hat{\rho}_a(\hat{g})|$. We will now attempt to study the convergence in probability properties of this term:

$$|\hat{\rho}_a(g) - \hat{\rho}_a(\hat{g})| = \left| \frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} \mathbb{I}(c_i \geq \alpha_a) - \frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} \mathbb{I}\left(c_i \geq \hat{\alpha}_a^+\right) \right| \leq \frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} Z_i,$$

where $Z_i = |\mathbb{I}(c_i \geq \alpha_a) - \mathbb{I}(c_i \geq \hat{\alpha}_a^+)| = \mathbb{I}(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})$.

We will now show that the numerator converges to zero. Note that conditioned on D_α, $\hat{\alpha}_a^+$ is fixed. Since D_α and D_β are sampled independently, $\{Z_i, i \in [D_{\beta,a}]\}$ are iid and bounded. Hence, by the weak law of large numbers, $\frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} Z_i \xrightarrow{p} \mathbb{E}_{D_{\beta,a}}[Z_i | D_\alpha, A = a] = \mathbb{E}_{D_{\beta,a}}[\mathbb{I}(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})) | D_\alpha, A = a] = \Pr(c \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\}) | D_\alpha, A = a) = |F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|$, where F_a is the group-conditional CDF of $c = \eta(x, a)$, and the absolute difference accounts for both scenarios ($\hat{\alpha}_a^+ > \alpha_a$ and $\hat{\alpha}_a^+ \leq \alpha_a$).

We are now ready to apply the Continuous Mapping Theorem, since F_a is continuous by Assumption 2.1. Since we ensure $\alpha_a \in (0, 1)$, by CMT, $F_a(\hat{\alpha}_a^+) \xrightarrow{p} F_a(\alpha_a)$, since $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ (Theorem 4.1), and therefore $\frac{1}{N_{\beta,a}} \sum_{i \in [N_{\beta,a}]} Z_i \xrightarrow{p} 0$ and:

$$|\hat{\rho}_a(g) - \hat{\rho}_a(\hat{g})| \leq \frac{1}{N_{\beta,a}} \sum_{i \in [N_{\beta,a}]} Z_i \xrightarrow{p} |F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)| \xrightarrow{p} 0.$$

With this, we now have that all our components of $|\Delta(g) - \hat{\Delta}(\hat{g})|$ converge to zero in probability and hence we can claim that $\Pr\left(|\hat{\Delta}_{DP}(\hat{g}) - \Delta_{DP}(g)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$.

(2) **Bias**: We now want to bound the bias of our estimator, $|\Delta_{DP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]|$.

$$\begin{aligned}
&|\Delta_{DP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]| \\
&= |\mathbb{E}_{D_\beta, D_\alpha}\left[\Delta_{DP}(g) - \hat{\Delta}_{DP}(\hat{g})\right]| \leq \mathbb{E}_{D_\beta, D_\alpha}\left[|\hat{\Delta}_{DP}(\hat{g}) - \Delta_{DP}(g)|\right] \\
&\leq \mathbb{E}_{D_\beta, D_\alpha}[|\hat{\rho}_1(\hat{g}) - \hat{\rho}_0(\hat{g}) - \rho_1(g) + \rho_0(g)|] \\
&\text{(By using the Reverse Triangle Inequality)} \\
&\leq \mathbb{E}_{D_{\beta,1}, D_\alpha}[|\rho_1(g) - \hat{\rho}_1(g)|] + \mathbb{E}_{D_{\beta,0}, D_\alpha}[|\rho_0(g) - \hat{\rho}_0(g)|] \\
&+ \mathbb{E}_{D_{\beta,1}, D_\alpha}[|\hat{\rho}_1(g) - \hat{\rho}_1(\hat{g})|] + \mathbb{E}_{D_{\beta,0}, D_\alpha}[|\hat{\rho}_0(g) - \hat{\rho}_0(\hat{g})|] \\
&\text{(By introducing an intermediate estimator } \hat{\rho}_a(g), \\
&\text{and noting that the expectation is over group-specific samples } D_{\beta.a}).
\end{aligned}$$

Once again, we will only focus on deriving the statistical bias for an arbitrary group $A = a$. Consider the first term. At this point, note that $\hat{\rho}_a(g)$ is an unbiased estimate of $\rho_a(g)$, and hence by Lemma B.2,

$$\mathbb{E}_{D_{\beta,a}, D_\alpha}[|\hat{\rho}_a(g) - \rho_a(g)|] \leq \sqrt{Var(\hat{\rho}_a(g))} \leq \frac{1}{2\sqrt{N_{\beta,a}}},$$

where the last step follows due to Popoviciu's Inequality on $\mathbb{I}(c_i \geq \alpha_a)$.

We can now focus on the second term: $\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\hat{\rho}_a(g)-\hat{\rho}_a(\hat{g})|\right]$.

$$\begin{aligned}\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\hat{\rho}_a(\hat{g})-\hat{\rho}_a(g)|\right] &= \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\left|\frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}\left(\mathbb{I}\left(c_i\geq\hat{\alpha}_a^+\right)-\mathbb{I}\left(c_i\geq\alpha_a\right)\right)\right|\right]\\ &\leq \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}\left|\mathbb{I}\left(c_i\geq\hat{\alpha}_a^+\right)-\mathbb{I}\left(c_i\geq\alpha_a\right)\right|\right]\end{aligned}$$

From our consistency analysis, we know that $Z_i = \left|\mathbb{I}\left(c_i\geq\alpha_a\right)-\mathbb{I}\left(c_i\geq\hat{\alpha}_a^+\right)\right| = \mathbb{I}\left(c_i\in(\min\{\hat{\alpha}_a^+,\alpha_A\},\max\{\hat{\alpha}_a^+,\alpha_a\})\right)$, and hence:

$$\begin{aligned}&\mathbb{E}_{D_{\beta,a},D_\alpha}\left[\frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}Z_i\right]\\ &=\frac{1}{N_{\beta,a}}\sum_{i\in[D_{\beta,a}]}\mathbb{E}_{D_{\beta,a},D_\alpha}\left[Z_i\right]=\mathbb{E}_{D_{\beta,a},D_\alpha}\left[\mathbb{I}\left(c_i\in(\min\{\hat{\alpha}_a^+,\alpha_a\},\max\{\hat{\alpha}_a^+,\alpha_a\})\right)\right]\\ &=\mathbb{E}_{D_\alpha}\left[\mathbb{E}_{D_{\beta,a}}\left[\mathbb{I}\left(c_i\in(\min\{\hat{\alpha}_a^+,\alpha_a\},\max\{\hat{\alpha}_a^+,\alpha_a\})\right)\right]|D_\alpha\right]=\mathbb{E}_{D_\alpha}\left[|F_a(\alpha_a)-F_a(\hat{\alpha}_a^+)|\right].\end{aligned}$$

For a fixed $\gamma\in(0,1)$, let $\mathcal{G}=\{|\hat{\alpha}_a^+-\alpha_a|\leq\gamma_a\}$ be the good event that the threshold error is at most γ_a. Then,

$$\begin{aligned}&\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\hat{\rho}_a(\hat{g})-\hat{\rho}_a(g)|\right]\leq\mathbb{E}_{D_\alpha}\left[|F_a(\alpha_a)-F_a(\hat{\alpha}_a^+)|\right]\\ &=\mathbb{E}_{D_\alpha}\left[|F_a(\alpha_a)-F_a(\hat{\alpha}_a^+)|\mathbb{1}_{\mathcal{G}}\right]+\mathbb{E}_{D_\alpha}\left[|F_a(\alpha_a)-F_a(\hat{\alpha}_a^+)|\mathbb{1}_{\mathcal{G}^C}\right]\\ &\leq L_a\gamma_a+\Pr\left(\mathcal{G}^c\right)\text{ (Using Assumption 2.2 and using }|F_a(\alpha_a)-F_a(\hat{\alpha}_a^+)|\leq 1)\\ &=L_a\gamma_a+\Pr\left(|\alpha_a-\hat{\alpha}_a^+|>\gamma_a\right).\end{aligned}$$

Overall,

$$|\Delta_{DP}(g)-\mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]|\leq\sum_{a\in\{0,1\}}\left(\frac{1}{2\sqrt{N_{\beta,a}}}+L_a\gamma_a+\Pr\left(|\alpha_a-\hat{\alpha}_a^+|>\gamma_a\right)\right).$$

Using the sample complexity bound from Theorem 4.1 and the dominant term, $\gamma_a=\frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$, giving us:

$$|\Delta_{DP}(g)-\mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]|\leq\sum_{a\in\{0,1\}}\left(\frac{1}{2\sqrt{N_{\beta,a}}}+\frac{L_a|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}+\delta_a\right).$$

Setting $N_\beta=\min\{N_{\beta,0},N_{\beta,1}\}$, we can obtain:

$$|\Delta_{DP}(g)-\mathbb{E}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{DP}(\hat{g})\right]|=\mathcal{O}\left(\frac{1}{\sqrt{N_\beta}}+|\lambda|\cdot\max\{L_0,L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right).$$

(3) **Sample Complexity**: To bound the sample complexity, we need to bound the following event:

$$\Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(g)| > \epsilon\right) \leq \delta.$$

Let $\Delta(\hat{g})$ denote the true DP difference with respect to the estimated threshold $\hat{\alpha}_a^+$. Consider the following quantity and its breakdown via the triangle inequality:

$$|\hat{\Delta}(\hat{g}) - \Delta(g)| \leq |\hat{\Delta}(\hat{g}) - \Delta(\hat{g})| + |\Delta(\hat{g}) - \Delta(g)| \text{ (By Triangle Inequality)}$$

By the union bound, we can upper bound our total error

$$\Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(g)| > \epsilon\right) \leq \Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(\hat{g})| > \epsilon/2\right) + \Pr\left(|\Delta(\hat{g}) - \Delta(g)| > \epsilon/2\right).$$

We can first use the Hoeffding bound for the first term by noting that $|\hat{\Delta}(\hat{g}) - \Delta(\hat{g})| \leq \sum_{a\in\mathcal{A}} |\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})|$:

$$\Pr\left(|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})| > \epsilon/4\right) \leq 2\exp\left(\frac{-2N_{\beta,a}\epsilon^2}{16}\right) \leq \frac{\delta}{4} \implies N_{\beta,a} \geq \frac{8\log(\frac{8}{\delta})}{\epsilon^2}.$$

For both groups, we would need $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log(\frac{1}{\delta})\right)$.

Next, consider the second term:

$$\begin{aligned}|\Delta(\hat{g}) - \Delta(g)| &\leq \sum_{a\in\mathcal{A}} |\rho_a(g) - \rho_a(\hat{g})| = \sum_{a\in\mathcal{A}} |\Pr\left(C \geq \alpha_a | A = a\right) - \Pr\left(C \geq \hat{\alpha}_a^+ | A = a\right))| \\ &= \sum_{a\in\mathcal{A}} |1 - F_a(\alpha_a) - 1 + F_a(\hat{\alpha}_a^+)| \leq \sum_{a\in\mathcal{A}} L_a \cdot |\alpha_a - \hat{\alpha}_a^+| \text{ (Using Assumption 2.2).}\end{aligned}$$

Now, we can figure out the sample complexity of each term:

$$\Pr\left(|\rho_a(g) - \rho_a(\hat{g})| > \epsilon/4\right) = \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{4L_a}\right) < \frac{\delta}{4}.$$

Now, we can use Theorem 4.1 to get the number of samples required to upper bound the above error: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \cdot \max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$. □

Theorem D.7. *(Proof of the EOP part of Theorem 4.6 in the main text) Suppose a dataset D_β is used to construct $\hat{\Delta}_{EOP}(g_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the EOP threshold (Theorem 4.1). Then the following holds:*

1. *Consistency:* $\hat{\Delta}_{EOP}(g_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{EOP}(g_\alpha)$.
2. *Bias:* $|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda|\cdot\max\{L_0, L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}}\right)$.
3. *Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{g}) - \Delta(g)| \leq \epsilon$, as long as $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$ and $N_\beta = \mathcal{O}\left\{\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right\}$.*

Proof. For the sake of this proof, we will denote $\Delta_{EOP}(g_\alpha)$ as $\Delta(g)$ and $\hat{\Delta}_{EOP}(g_{\hat{\alpha}_a^+})$ as $\hat{\Delta}(\hat{g})$. Then an intermediate estimator $\hat{\Delta}_{EOP}(g_{\alpha_a})$ can be denoted as $\hat{\Delta}(g)$. We similarly overload the notations in this fashion for the positive rate estimator $\rho_{\mathrm{TPR},a}$ with ρ_a.

Recalling the definitions of our estimations,

$$\rho_{\mathrm{TPR},a}(g_\alpha) = \Pr\left(g_\alpha = 1 | Y = 1, A = a\right) = \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(\eta(x,a) \geq \alpha_a\right) \cdot \Pr\left(Y = 1 | X = x, A = a\right)\right]}{\pi_{1|a}}$$
$$= \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \alpha_a\right) \cdot C\right]}{\pi_{1|a}},$$

and,

$$\hat{\rho}_{\mathrm{TPR},a}(g_{\hat{\alpha}_a^+}) = \frac{\frac{1}{N_{\beta,a}} \sum_{i \in [D_{\beta,a}]} (c_i(\mathbb{I}\left(c_i \geq \hat{\alpha}_a^+\right))}{\hat{\pi}_{1|a}^+} = \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(c \geq \hat{\alpha}_a^+\right) c]}{\hat{\pi}_{1|a}^+}$$

(1) **Consistency**: Most of the inequality arguments in the proof are a direct consequence of the triangle inequality, by introducing an intermediate estimate of the TPR ($\hat{\rho}_a(g)$).

We want to show that $\Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(g)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$. By Slutsky's theorem, it will be enough to upper-bound our target event as a sum of simpler events and show that if each of them converges in probability to zero, then we can claim our final result.

Consider the absolute difference term and apply the reverse triangle inequality ($||A| - |B|| \leq |A - B|$):

$$|\Delta(g) - \hat{\Delta}(\hat{g})| \leq |\rho_1(g) - \rho_0(g) - \hat{\rho}_1(\hat{g}) + \hat{\rho}_0(\hat{g})|$$
$$\leq |\rho_1(g) - \rho_1(\hat{g})| + |\rho_0(g) - \rho_0(\hat{g})| + |\rho_1(\hat{g}) - \hat{\rho}_1(\hat{g})| + |\rho_0(\hat{g}) - \hat{\rho}_0(\hat{g})|.$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$.

Consider $|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})|$. $\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(c \geq \hat{\alpha}_a^+\right) c] \to \mathrm{E}_{X|A=a}[\mathbb{I}\left(c \geq \hat{\alpha}_a^+\right) c]$ and $\hat{\pi}_{1|a}^+ \xrightarrow{p} \pi_{1|a}$ due to the weak law of large numbers, since conditioned on D_α, $\hat{\alpha}_a^+$ and $\hat{\pi}_{1|a}^+$ are fixed numbers.

Since $\pi_{1|a} > 0$ by assumption, by Slutsky's theorem, $\hat{\rho}_a(\hat{g}) = \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(c \geq \hat{\alpha}_a^+\right) c]}{\hat{\pi}_{1|a}^+} \xrightarrow{p} \rho_a(\hat{g})$.

Next, consider the term $|\rho_a(\hat{g}) - \rho_a(g)|$. Since the denominator for both is a constant $\pi_{1|a}$, we will focus on showing that the difference in the numerator converges to zero.

$$|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \alpha_a\right) C\right] - \mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right]| \leq \mathbb{E}_{X|A=a}\left[Z\right]$$

where $Z = |\mathbb{I}\left(C \geq \alpha_a\right) - \mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)| = \mathbb{I}(C \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\}))$ and $C \leq 1$. Using this we can write $\mathbb{E}_{X|A=a}\left[Z\right] = \Pr\left(\min\{\hat{\alpha}_a^+, \alpha_a\} \leq C < \max\{\hat{\alpha}_a^+, \alpha_a\} | A = a\right) = |F_a(\hat{\alpha}_a^+) - F_a(\alpha_a)|$. Since F_a is continuous, and from Theorem 4.1, $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$, by the continuous mapping theorem, $F_a(\hat{\alpha}_a^+) \xrightarrow{p} F_a(\alpha_a)$. Combining all the pieces, $\hat{\Delta}_{EOP}(g_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{EOP}(g_\alpha)$.

(2) **Bias**: We now want to bound the bias of our estimator, $|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]|$.

$$|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]| = |\mathbb{E}_{D_\beta, D_\alpha}\left[\Delta_{EOP}(g) - \hat{\Delta}_{EOP}(\hat{g})\right]|$$
$$\leq \mathbb{E}_{D_\beta, D_\alpha}\left[|\hat{\Delta}_{EOP}(\hat{g}) - \Delta_{EOP}(g)|\right] \leq \mathbb{E}_{D_\beta, D_\alpha}\left[|\hat{\rho}_1(\hat{g}) - \hat{\rho}_0(\hat{g}) - \rho_1(g) + \rho_0(g)|\right]$$
(By using the Reverse Triangle Inequality)
$$\leq \mathbb{E}_{D_{\beta,1}, D_\alpha}\left[|\rho_1(g) - \rho_1(\hat{g})|\right] + \mathbb{E}_{D_{\beta,0}, D_\alpha}\left[|\rho_0(g) - \rho_0(\hat{g})|\right]$$
$$+ \mathbb{E}_{D_{\beta,1}, D_\alpha}\left[|\rho_1(\hat{g}) - \hat{\rho}_1(\hat{g})|\right] + \mathbb{E}_{D_{\beta,0}, D_\alpha}\left[|\rho_0(\hat{g}) - \hat{\rho}_0(\hat{g})|\right]$$
(By introducing an intermediate estimator $\rho_a(\hat{g})$,
and noting that the expectation is over group-specific samples $D_{\beta.a}$).

Once again, we will only focus on deriving the statistical bias for an arbitrary group $A = a$. Consider the term $\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \rho_a(g)|\right]$. Since the denominator for both is a constant $\pi_{1|a}$, we will focus on showing that the difference in the numerator converges to zero. Reusing our analysis from the previous proof on consistency and using Assumption 2.2, we can write:

$$\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \rho_a(g)|\right] \leq \frac{1}{\pi_{1|a}}\mathbb{E}_{D_\alpha}\left[|F_a(\hat{\alpha}_a^+) - F_a(\alpha_a)|\right] \leq \frac{L_a}{\pi_{1|a}}\mathbb{E}_{D_\alpha}\left[|\hat{\alpha}_a^+ - \alpha_a|\right]$$

From Theorem 4.1, we know that with $1-\delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$. We also know that $|\alpha_a - \hat{\alpha}_a^+| \leq 1$. Using this, we get:

$$\mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] = \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon_a\right)\right] + \mathbb{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon_a\right)\right]$$
$$\leq \epsilon_a \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon_a\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geq \epsilon_a\right) \text{ (Since } |\alpha_a - \hat{\alpha}_a^+| \leq 1) \leq \epsilon_a + \delta_a.$$

Using $\epsilon_a = \frac{|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}}$, our final bound becomes:

$$\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \rho_a(g)|\right] \leq \frac{L_a}{\pi_{1|a}}\left(\frac{|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}} + \delta_a\right)$$

Next, consider the term $\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})|\right]$.

$$\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})|\right] = \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\left|\frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}\right|\right]$$
$$= \mathrm{E}_{D_{\beta,a},D_\alpha}\Bigg[\Bigg|\frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]}{\pi_{1|a}}$$
$$+ \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}\Bigg|\Bigg]$$
(By introducing an intermediate term $\frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]}{\pi_{1|a}}$)
$$\leq \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\frac{1}{\pi_{1|a}} \cdot |\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]|\right]$$
$$+ \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C] \cdot \left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right|\right]$$
$$\leq \frac{1}{\pi_{1|a}}\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]|\right]$$
$$+ \mathbb{E}_{D_{\beta,a},D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right|\right] \text{ (Since } \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C] \leq 1)$$

Conditioned on D_α, $\hat{\alpha}_a^+$ is fixed. Therefore, we can use Lemma B.2 to upper bound the term $\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]|\right]$:

$$
\begin{aligned}
&\mathbb{E}_{D_{\beta,a},D_\alpha}\left[\left|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]\right|\right] \\
&= \mathbb{E}_{D_{\beta,a}}\left[\left|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]\right| | D_\alpha\right] \\
&= \sqrt{Var(\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C] | D_\alpha)} = \sqrt{\frac{1}{N_{\beta,a}} Var(\mathbb{I}\left(C \geq \alpha_a^+\right) C)} \\
&\leq \frac{1}{2\sqrt{N_{\beta,a}}} \text{ (Using Popoviciu's Inequality).}
\end{aligned}
$$

To bound $\mathbb{E}_{D_{\beta,a},D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right|\right] = \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right|\right]$ (Since $\hat{\pi}^+_{1|a}$ is estimated using D_α), let $\mathcal{G} = \{\hat{\pi}_{1|a} \geq \frac{\pi_{1|a}}{2}\}$ denote the event that $\hat{\pi}_{1|a}$ and $\pi_{1|a}$ are close. Note that on $\mathcal{G}$, $\hat{\pi}^+_{1|a} = \hat{\pi}_{1|a}$ for sufficiently many samples ($1/N_{\alpha,a} \leq \pi_{1|a}/2$):

$$
\begin{aligned}
&\mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi^+_{1|a}} - \frac{1}{\hat{\pi}_{1|a}}\right|\right] = \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}}\right| \mathbb{1}_{\mathcal{G}}\right] + \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right| \mathbb{1}_{\mathcal{G}_c}\right] \\
&= \mathbb{E}_{D_\alpha}\left[\frac{|\hat{\pi}_{1|a} - \pi_{1|a}|}{\hat{\pi}_{1|a}\pi_{1|a}} \mathbb{1}_{\mathcal{G}}\right] + \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right| \mathbb{1}_{\mathcal{G}_c}\right] \\
&\leq \frac{2}{\pi^2_{1|a}} \mathbb{E}_{D_\alpha}\left[|\hat{\pi}_{1|a} - \pi_{1|a}|\right] + \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right| \mathbb{1}_{\mathcal{G}_c}\right] \text{ (On } \mathcal{G}, \hat{\pi}_{1|a} \geq \pi_{1|a}/2) \\
&\leq \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \mathbb{E}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}^+_{1|a}}\right| \mathbb{1}_{\mathcal{G}_c}\right] \text{ (Using Lemma B.2 and Popoviciu's Inequality)} \\
&\leq \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \mathbb{E}_{D_\alpha}\left[\left(\frac{1}{\pi_{1|a}} + N_{\alpha,a}\right) \mathbb{1}_{\mathcal{G}_c}\right] \text{ (Using the form of our truncated estimator } \hat{\pi}^+_{1|a}) \\
&\leq \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}} + N_{\alpha,a}\right) \Pr\left(\mathcal{G}_c\right) \text{ (Using the Cauchy-Schwartz Inequality)} \\
&\leq \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}} + N_{\alpha,a}\right) \exp\left(-\frac{N_{\alpha,a}\pi^2_{1|a}}{2}\right) \text{ (Using the Hoeffding's Inequality)}
\end{aligned}
$$

Our upper bound becomes:

$$
\mathbb{E}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})|\right] \leq \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}} + N_{\alpha,a}\right) \exp\left(-\frac{N_{\alpha,a}\pi^2_{1|a}}{2}\right)
$$

and our overall upper bound on the bias of the estimator takes the form:

$$
\begin{aligned}
&|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]| \\
&\leq \sum_{a\in\{0,1\}} \left(\frac{L_a}{\pi_{1|a}} \left(\frac{|\lambda|}{\tau^2 \pi_{1a}^2} \sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}} + \delta_a \right) + \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi_{1|a}^2 \sqrt{N_{\alpha,a}}} \right. \\
&\left. + \left(\frac{1}{\pi_{1|a}} + N_{\alpha,a}\right) \exp\left(-\frac{N_{\alpha,a}\pi_{1|a}^2}{2}\right) \right) \\
&= \sum_{a\in\{0,1\}} \mathcal{O}\left(\frac{L_a}{\pi_{1|a}} \left(\frac{|\lambda|}{\tau^2 \pi_{1a}^2} \sqrt{\frac{\log(2/\delta_a)}{N_\alpha}} + \delta_a \right) + \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi_{1|a}^2 \sqrt{N_{\alpha,a}}} \right)
\end{aligned}
$$

Setting $\delta = \min\{\delta_0, \delta_1\}$, $N_\alpha = \min\{N_{\alpha,0}, N_{\alpha,1}\}$ and $N_\beta = \min\{N_{\beta,0}, N_{\beta,1}\}$:

$$
|\Delta_{EOP}(g) - \mathbb{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EOP}(\hat{g})\right]| = \mathcal{O}\left(\frac{|\lambda| \cdot \max\{L_0, L_1\}}{\tau^2} \sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}} \right)
$$

(3) **Sample Complexity**: To bound the sample complexity, we need to bound the following event:

$$
\Pr\left(|\hat{\Delta}(\hat{g}) - \Delta(g)| > \epsilon \right) \leq \delta.
$$

Let $\Delta(\hat{g})$ denote the true EOP difference with respect to the estimated threshold $\hat{\alpha}_a^+$. Consider the absolute difference term and apply the reverse triangle inequality ($||A| - |B|| \leq |A - B|$):

$$
\begin{aligned}
&|\Delta(g) - \hat{\Delta}(\hat{g})| \leq |\rho_1(g) - \rho_0(g) - \hat{\rho}_1(\hat{g}) + \hat{\rho}_0(\hat{g})| \\
&\leq |\rho_1(g) - \rho_1(\hat{g})| + |\rho_0(g) - \rho_0(\hat{g})| + |\rho_1(\hat{g}) - \hat{\rho}_1(\hat{g})| + |\rho_0(\hat{g}) - \hat{\rho}_0(\hat{g})|.
\end{aligned}
$$

By the union bound, we want to ensure the following:

$$
\Pr\left(|\Delta(g) - \hat{\Delta}(\hat{g})| > \epsilon \right) \leq \sum_{a\in\{0,1\}} \Pr\left(|\rho_a(g) - \rho_a(\hat{g})| > \epsilon/4 \right) + \Pr\left(|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})| > \epsilon/4 \right).
$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$. From our analysis in previous part, we know that $|\rho_a(g) - \rho_a(\hat{g})| \leq \frac{L_a}{\pi_{1|a}} |\hat{\alpha}_a^+ - \alpha_a|$ and therefore $\Pr\left(|\rho_a(g) - \rho_a(\hat{g})| > \epsilon/4 \right) \leq \Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \frac{\epsilon \pi_{1|a}}{4L_a} \right)$.

Using Theorem 4.1, recall that with $1 - \delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geq \frac{1}{2t^2}\log(\frac{2}{\delta})$ for $t = \min\left\{ \frac{\pi_{1a}}{2}, \frac{\tau \pi_{1a}^2}{4|\lambda|}, \frac{\epsilon \tau^2 \pi_{1a}^2}{4|\lambda|} \right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon \right) \leq \delta$. Using these bounds with our new ϵ and δ we get:

$$
\begin{aligned}
\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon \pi_{1|a}}{4L_a} \right) \leq \delta/4 \implies N_\alpha &\geq \frac{1}{2}\log\left(\frac{8}{\delta}\right) \max\left\{ \frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau \pi_{1a}^2}, \frac{16|\lambda| L_a}{\epsilon \tau^2 \pi_{1|a} \pi_{1a}^2} \right\}^2 \\
&= \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2 \tau^4} \log\left(\frac{1}{\delta}\right) \right).
\end{aligned}
$$

Now we focus on the other event. Just like the bias proof, we will proceed by introducing an intermediate estimator:

$$
\begin{aligned}
|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})| &= \left| \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]}{\hat{\pi}_{1|a}^+} \right| \\
&= \left| \frac{\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]}{\pi_{1|a}} \right. \\
&\left. + \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]}{\hat{\pi}_{1|a}^+} \right| \\
&\text{(By introducing an intermediate term } \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]}{\pi_{1|a}}\text{)} \\
&\leq \frac{1}{\pi_{1|a}} \cdot |\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]| \\
&+ \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C] \cdot \left| \frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+} \right| \\
&\leq \frac{1}{\pi_{1|a}} |\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]| + \left| \frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+} \right| \\
&\text{(Since } \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C] \leq 1\text{)}
\end{aligned}
$$

Now we can use the union bound:

$$
\begin{aligned}
&\Pr\left(|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})| > \epsilon/4\right) \\
&\leq \Pr\left(\frac{1}{\pi_{1|a}} |\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right) C]| > \epsilon/8 \right) \\
&+ \Pr\left(\left| \frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+} \right| > \epsilon/8 \right)
\end{aligned}
$$

To further simplify $\left| \frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+} \right|$, we can borrow our analysis from the bias proof. Recall $\mathcal{G} = \{\hat{\pi}_{1|a} \geq \frac{\pi_{1|a}}{2}\}$ to be the event that $\hat{\pi}_{1|a}$ and $\pi_{1|a}$ are close. Note that on $\mathcal{G}$, $\hat{\pi}_{1|a}^+ = \hat{\pi}_{1|a}$ for sufficiently many samples ($1/N_{\alpha,a} \leq \pi_{1|a}/2$). On this good event $\mathcal{G}$, we derived that $\left| \frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+} \right| \leq \frac{2}{\pi_{1|a}^2} |\hat{\pi}_{1|a} - \pi_{1|a}|$.

$$
\begin{aligned}
&\Pr\left(|\rho_a(\hat{g}) - \hat{\rho}_a(\hat{g})| > \epsilon/4\right) \\
&\leq \Pr\left(\frac{1}{\pi_{1|a}}|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]| > \epsilon/8\right) \\
&+ \Pr\left(\Big|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\Big| > \epsilon/8\right) \\
&= \Pr\left(\frac{1}{\pi_{1|a}}|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]| > \epsilon/8\right) \\
&+ \Pr\left(\Big|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\Big| > \epsilon/8, \mathcal{G}\right) + \Pr\left(\Big|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\Big| > \epsilon/8, \mathcal{G}_C\right) \\
&\leq \Pr\left(\frac{1}{\pi_{1|a}}|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]| > \epsilon/8\right) \\
&+ \Pr\left(\frac{2}{\pi_{1|a}^2}|\hat{\pi}_{1|a} - \pi_{1|a}| > \epsilon/8\right) + \Pr\left(\mathcal{G}_C\right) \\
&= \Pr\left(|\mathbb{E}_{X|A=a}\left[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geq \hat{\alpha}_a^+\right)C]| > \frac{\pi_{1|a}\epsilon}{8}\right) \\
&+ \Pr\left(|\hat{\pi}_{1|a} - \pi_{1|a}| > \frac{\pi_{1|a}^2\epsilon}{16}\right) + \Pr\left(\hat{\pi}_{1|a} < \frac{\pi_{1|a}}{2}\right) \\
&\leq 2\left(\exp\Big(-\frac{N_{\beta,a}\pi_{1|a}^2\epsilon^2}{32}\Big) + \exp\Big(-\frac{N_{\alpha,a}\pi_{1|a}^4\epsilon^2}{128}\Big) + \exp\Big(-\frac{N_{\alpha,a}\pi_{1|a}^2}{2}\Big)\right)
\end{aligned}
$$

(By using Hoeffding's Inequality on each event).

We can now derive our sample complexities by splitting out $\delta/4$ budget between all three terms to obtain $N_{\beta,a} \geq \frac{32}{\pi_{1|a}^2\epsilon^2}\log\left(\frac{24}{\delta}\right)$ and $N_{\alpha,a} \geq \max\left\{\frac{2}{\pi_{1|a}^2}\log\left(\frac{24}{\delta}\right), \frac{128}{\pi_{1|a}^4\epsilon^2}\log\left(\frac{24}{\delta}\right)\right\}$.

Overall, our sample complexities take the following form:

$$
N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right) \text{ and } N_\beta = \mathcal{O}\left\{\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right\}.
$$

□

E Details for the Tradeoff Estimation algorithm and Experiments

With all the tools in place, we now present the algorithm for estimating the Pareto frontier using the estimators derived using soft labels in the previous section. We will now describe the algorithm formally in Algorithm 1 and discuss each design choice in more detail.

1-2 Assuming access to sufficiently large, disjoint datasets $D_\alpha.D_\beta$ sampled i.i.d. from our target distribution, we obtain estimates for all the distribution-dependent quantities $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}_a, \hat{\pi}_{ya}, \hat{\pi}_{y|a}\}$. Using these priors $\hat{\Pi}$ and any λ, we can estimate our thresholds $\hat{\alpha}_a^+$ (AlphaEst) and then our risk R and unfairness Δ (ErrUnfEst).

3-4 If the Bayes optimal classifier (at $\lambda = 0$) already satisfies the required fairness ($|\hat{\Delta}_0| \leq \gamma$), then we are done.

5 Depending on sign of $\hat{\Delta}_0$, we can set $\lambda_{\max}$ from Lemma 3.8 to fix our range for the bisection search.

Algorithm 1 Tradeoff Estimation with Soft Labels

1: **Input:** Disjoint datasets D_α, D_β; Unfairness Tolerance γ, Precision ξ; Priors $\hat{\Pi}$
2: **Functions:**
- $\mathrm{AlphaEst}(\lambda, \hat{\Pi}) \rightarrow \hat{\alpha}_a^+$ (Function to estimate the cost-sensitive thresholds α_a given a λ and distribution priors);
- $\mathrm{ErrUnfEst}(\lambda, \mathrm{AlphaEst}(\lambda, \hat{\Pi})) \rightarrow R(g_{\hat{\alpha}_a^+}), \Delta(g_{\hat{\alpha}_a^+})$ (Function to estimate the error and unfairness, given a threshold α_a computed using the previous function)
3: $\hat{R}_0, \hat{\Delta}_0 \leftarrow \mathrm{ErrUnfEst}(0, \hat{\alpha}_0)$
4: **if** $|\hat{\Delta}_0| \leq \gamma$ **then** return No-Tradeoff.
5: Set range: **if** $\hat{\Delta}_0 > 0$ **then** $\lambda \in [0, \frac{1}{\inf w}]$ **else** $\lambda \in [-\frac{1}{\inf w}, 0]$
6: points $\leftarrow \{\}$
7: **while** $\lambda_{\max} - \lambda_{\min} > \xi$ **do**
8: $\quad \lambda \leftarrow (\lambda_{\max} + \lambda_{\min})/2$
9: $\quad \hat{R}, \hat{\Delta} \leftarrow \mathrm{ErrUnfEst}(\lambda, \mathrm{AlphaEst}(\lambda, \hat{\Pi}))$
10: $\quad$ Append $(\lambda, \hat{R}, \hat{\Delta})$ to points
11: $\quad$ **if** $|\hat{\Delta}| \leq \gamma$ **then** break loop
12: $\quad$ **if** $\hat{\Delta} > \gamma$ **then** $\lambda_{\max} \leftarrow \lambda$ **else** $\lambda_{\min} \leftarrow \lambda$
13: **end while**
14: return points

8-12 We compute λ and calculate the updated $\hat{R}, \hat{\Delta}$. If we immediately achieve our required unfairness γ, we are done. Otherwise, depending on the sign, we update our bound $\lambda_{\max}/\lambda_{\min}$ and continue until the gap between $\lambda_{\min}$ and $\lambda_{\max}$ is under a fixed precision ξ.

Details of Data and Experiments: The synthetic dataset setup in our experiments is derived from Zeng et al. [12], which consists of a binary label $y \in \{0, 1\}$ and a binary sensitive attribute $a \in \{0, 1\}$. We sample data from subgroup-specific 100-dimensional Gaussian distributions $X \sim \mathcal{N}(\mu_{ya}, \sigma^2 I_{100})$ where I_{100} denotes a 100×100 identity matrix and all entries of μ_{ya} are sampled from $U(0, 1)$. We sample $20,000$ training examples and 5000 test examples.

The overall setup of soft labels is borrowed from previous state-of-the-art studies in Bayes error estimation: Ishida et al. [6] and Ushio et al. [14]. CIFAR-10H [49] is a dataset constructed using 50 additional annotations of the CIFAR-10 test dataset. The ground-truth soft label is constructed as the average of the underlying hard labels. Since the original dataset is multi-class (10 classes), the binary labels are constructed by grouping classes into two: land vs. other animals, animal vs. artifacts, odd vs. even, and so on. Because we are working with unfairness constraints that also include a notion of a sensitive group, we assume one of these partitions is the class and the other is the sensitive group. Note that no dataset of soft labels exists for standard fairness datasets, and hence, we had to retrofit a popular benchmark from the soft-label-based Bayes error estimation literature. The code for setting up the CIFAR-10H experiments is from Ishida et al. [6] [6]. The code and data to reproduce the experiments is provided here.

[6] https://github.com/ishida-lab/irreducible

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main contributions have been well motivated and highlighted in the abstract and the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion section discusses limitations and future extensions of our work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation, while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions required are included in the main body and all the proofs are provided in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: A link to anonymized code to reproduce the empirical results has been included in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: A link to anonymized code to reproduce the empirical results has been included in the Appendix.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Data preparation scripts and the random seeds used are provided in the anonymized code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments are performed with bootstrapping, and the error bars are also reported with the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: Our experiments are lightweight and can be run on most systems with moderate compute power.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All points in the code of ethics have been conformed to. We do not believe our research process has caused harm — it does not involve human participation, and we use standard public datasets from the literature. Our research will help society curb discrimination by directly tackling the themes of bias and unfairness. We do not see any possible adverse impact on society due to our research work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the benefits of our work for practitioners and the society at large, and do not foresee any negative impact.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper does not involve any data or models that can cause harm or have a risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have credited the authors of the datasets and the codebases we used as baselines.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented, and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The link to the anonymized code implementation is included in the Appendix.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Human subjects not involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not used for any important, original or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.