AN OPTIMAL CRITERION FOR STEERING DATA DISTRIBUTIONS TO ACHIEVE EXACT FAIRNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

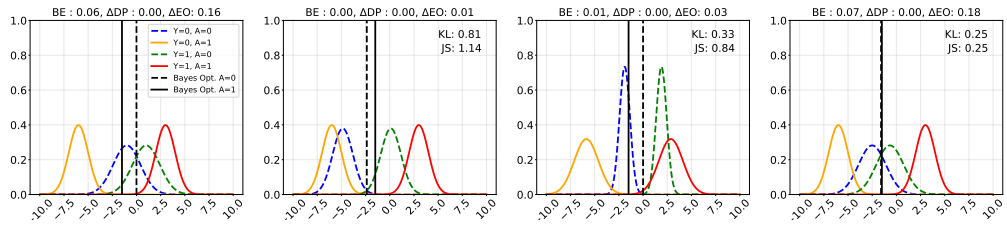
To fix the 'bias in, bias out' issue in fair machine learning, it is essential to get ideal training and validation data. Collecting ideal real-world data or generating ideal synthetic data requires a formal specification of *ideal* distribution that would guarantee fair outcomes by downstream models. Previous work on fair pre-processing does not address this gap, and could be significantly improved if it is resolved. We call a distribution as *ideal* distribution if the minimizer of any cost-sensitive risk on it is guaranteed to satisfy exact fairness (e.g., demographic parity, equal opportunity). Given any data distribution for fair classification, we formulate an optimization program to find its nearest *ideal* distribution in KL-divergence. This optimization is intractable as stated but we show how it can be solved efficiently when the distributions come from well-known parametric families (e.g., normal, log-normal). We empirically show on synthetic datasets that our ideal distributions are close to the given distributions and they can often suggest directions to steer the original distribution to improve both accuracy and fairness simultaneously.

1 INTRODUCTION

The importance of clean or ideal data in fair machine learning cannot be emphasized more. *Bias in, bias out* has been argued to be a root cause of unfair outcomes in machine learning models Buolamwini & Gebru (2018); Mayson (2019); Rambachan & Roth (2020); Cowgill et al. (2020). Models trained on biased data often learn, perpetuate, and amplify these biases. The problem of data bias is not about training data alone. Fair in-processing or fairness-constrained training on biased data cannot guarantee fairness on (unbiased) test data. Fair post-processing of model predictions using biased validation data cannot guarantee fairness on (unbiased) test data. Biased data used for assessment can lead to faulty fairness audits that may be hard to correct later in machine learning pipelines Biswas & Rajan (2021); Bakalar et al. (2021).

Popular fairness metrics are functions of both a given model and a data distribution, e.g., demographic parity (equal positive rates across demographic groups), equal opportunity (equal true positive rates across groups). They offer two natural ways to correct unfair outcomes: either by correcting the model or by correcting the data distribution. An *ideal* model can thus be defined as one that satisfies exact fairness, and fair in-processing tries to fit an *ideal* model to the given data distribution Agarwal et al. (2018); Donini et al. (2018). In this paper, we focus on the latter approach of finding an *ideal* data distribution instead. In that sense, we are closest to the fair pre-processing literature. A detailed discussion of pre-processing approaches is given in Appendix A.

A common goal of all fair pre-processing methods is to find an *ideal* data distribution close to the given distribution so that any downstream model trained on it must have guaranteed fairness. A stronger requirement that this should hold for downstream models optimized for multiple tasks leads to impossibility results Lechner et al. (2021). If all downstream classifiers are required to be fair, then the group-wise distributions must be nearly identical, which is absurd. Thus, we restrict our downstream models only to Bayes optimal classifiers for cost-sensitive risks. Our first contribution is to formally define an *ideal* distribution (Definition 3.1) as the one where the Bayes optimal classifier for any cost-sensitive risk satisfies exact fairness (e.g., satisfies equal opportunity perfectly). To operationalize this definition, we assume group and class-conditioned distributions come from well-known parametric families (e.g., Gaussian, log-normal) and show conditions on such ideal distributions (Propositions 3.2, 3.3). This allows us to convert the ideal distribution optimization



(a) Original Distribution (b) Ideal-Affirmative (c) Ideal-Changing All (d) Mean Matching

Figure 1: Comparison of Different Interventions for Changing Data Distributions for Exact Fairness. Figure (1a) captures the original distribution, its Bayes error (BE), and the unfairness differences (ΔDP and ΔEO). In Figure (1b), we only change the under-privileged group using Corollary C.2, and in Figure (1c) we change all four subgroups using Proposition C.3. Finally, in Figure (1d), we match the means of the two groups using Proposition C.4. Figures (1b) and Figures (1c) show that it is possible to construct 'ideal' distributions that are close to the given distribution where the Bayes Optimal classifier is maximally accurate and fair.

problem, which is generally intractable, to a tractable problem that can be solved efficiently in some cases and give closed form transformations (Theorem 4.1, Corollary C.2, Proposition C.3).

Bayes optimal classifier maximizes accuracy on a given distribution, and have been an important object of study in statistical machine learning Devroye et al. (1996). Fair Bayes optimal classifier maximizes accuracy subject to fairness constraints, and its mathematical characterization for binary fair classification has been important in fair classification Menon & Williamson (2018); Chzhen et al. (2019); Celis et al. (2021); Zeng et al. (2022). Blum & Stangl (2019) introduce a data bias model that injects under-representation and label bias in an original unbiased distribution to create biased data. They show that, for a stylized distribution under some conditions, the fair Bayes optimal classifier on the biased distribution recovers the Bayes optimal classifier on the original unbiased distribution. Their unbiased distribution is *ideal* by construction, i.e., the Bayes optimal classifier on their unbiased distribution is guaranteed to be perfectly fair. Sharma & Deshpande (2024) extend this observation to general hypothesis classes and distributions beyond the stylized setting of Blum & Stangl (2019). Blum et al. (2023) study fair Bayes optimal classifier whether its accuracy is robust to malicious corruptions in data distribution.

In contrast to these results, our focus is not on finding the *ideal* classifier but on finding the nearest *ideal* distribution. By definition, our *ideal* distribution has no trade-off between accuracy (or cost-sensitive utility) and fairness. If we find an *ideal* distribution close to our original distribution, we can steer our distribution towards reducing fairness-accuracy trade-off. Moreover, if the *ideal* distribution offers better accuracy, it suggests that we can steer our distribution to improve both accuracy and fairness simultaneously. We highlight this in Fig. 1, and later in Figures 2, 3, 4 and 5.

2 Problem Setup and Preliminaries

Let (X, A, Y) be a random data point from a joint distribution D over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, where $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ denote the sets of features, sensitive attributes, and class labels, respectively. For simplicity of exposition, we consider a binary labels ($\mathcal{Y} = \{0, 1\}$) and a binary sensitive attributes ($\mathcal{A} = \{0, 1\}$). Let $q_{ia} = \Pr(Y = i, A = a)$ and P_{ia} denote the distribution $X \mid Y = i, A = a$ with the probability density $p_{ia}(x) = \Pr\left(X = x \mid Y = i, A = a\right)$. When P_{ia}'s come from parametric families of distributions, we assume $\mathcal{X} = \mathbb{R}^d$. We work with the following well-known definitions of fairness in classficiation Dwork et al. (2012); Hardt et al. (2016); Barocas et al. (2019).

Definition 2.1. For the case of binary labels and sensitive attributes, a group-aware classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ satisfies: (1) *Demographic Parity* if the positive rates are equal across groups, i.e., $\Pr(h(X, A) = 1|A = 0) = \Pr(h(X, A) = 1|A = 1)$, and (2) *Equal Opportunity* if the true positive rates are equal across groups, i.e., $\Pr(h(X, A) = 1|Y = 1, A = 0) = \Pr(h(X, A) = 1|Y = 1, A = 1)$.

These lead to quantitative metrics of unfairness, e.g., $\Delta_{\text{DP}}(h, D)$ denotes the absolute value of difference between $\Pr(h(X, A) = 1|A = 0)$ and $\Pr(h(X, A) = 1|A = 1)$. Similarly, $\Delta_{\text{EO}}(h, D)$ denotes the absolute value of difference between $\Pr(h(X, A) = 1|Y = 1, A = 0)$

and $\Pr(h(X, A) = 1|Y = 1, A = 1)$. We consider group-aware classifiers. We are particularly interested in threshold classifiers $h_t(x, a)$ that apply a group and feature dependent threshold $t(x, a)$ to the class probability of an example: $h_t(x, a) = \mathbb{I}(\eta(x, a) \geq t(x, a))$ where $\eta(x, a) = \Pr(Y = 1|X = x, A = a)$. It is well-known that the Bayes optimal classifier for a given distribution has the form $t(x, a) = 1/2$ Devroye et al. (1996). For a cost matrix $C \in \mathbb{R}^{2\times 2}$ and the associated cost sensitive loss l_C, the Bayes optimal classifier is defined as $\mathbb{I}(\eta(x, a) \geq t_C)$, for a threshold $t_C = (c_{10} - c_{00})/(c_{10} - c_{00} + c_{01} - c_{11}) \in [0, 1]$, where c_{ij} denote the entries of the cost matrix $C \in \mathbb{R}^{2\times 2}$ Elkan (2001); Scott (2012); Koyejo et al. (2014); Singh & Khim (2022). We defer all proofs to appendix for a smoother flow of presentation.

3 Ideal Distributions for Fair Classification

We define a data distribution as *ideal* when minimizing any cost-sensitive risk on it is guaranteed to give exact fairness (e.g., demographic parity, equal opportunity). In practice, downstream models trained on a distribution are typically optimized for some performance or utility metric that may not be known in advance. Our definition of ideal distribution allows the flexibility to choose any cost-sensitive risk as the performance metric for downstream models and still gives exact fairness guarantee for any optimal model downstream.

Definition 3.1. Let $\mathcal{H}$ be a hypothesis class of group-aware classifiers $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ and let $\Delta(h, D)$ be a given unfairness metric, e.g., demographic parity difference, equal opportunity difference. Given a distribution D over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and a cost-sensitive risk defined by $C \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$, let $h^*_C = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \Pr(\ell_C(h(X, A), Y))$. We call D an *ideal distribution* if $\Delta(h^*_C, D) = 0$, for all $C \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$.

Examples of fairness metrics include Demographic Parity, Equal Opportunity and Equalized Odds (Definition 2.1) and examples of cost-sensitive risk include the usual $0 - 1$ loss and different performance metrics which are functions of the confusion matrix metrics Elkan (2001); Koyejo et al. (2014); Singh & Khim (2022). Our definition gets around the impossibility theorems about fair representation for multiple tasks Lechner et al. (2021). However, we need to be careful of two things. First, our definition should not be too restrictive to just force the group-conditioned distributions to be similar or identical, as that would be impractical. Second, we need an efficient and equivalent way of expressing the constraint of being ideal. We show how to express it as a parametric condition when the group and class-conditioned distributions belong to certain well-known parametric families of distributions. This helps in checking if a given distribution is ideal, and otherwise, finding its nearest ideal distribution.

3.1 Parametric Conditions for Ideal Distributions

Borrowing a simple set up of parametric distributions from previous work on fair machine learning Pierson et al. (2018), we assume that the class and group-conditioned feature distributions $X \mid Y = i, A = a$ belong to a parametric family of distributions, e.g., univariate or multivariate Gaussians, log-normal. In that case, we show that the property of being *ideal* (Definition 3.1) can be equivalently expressed as certain parametric conditions. For example, here is what we get when $X \mid Y = i, A = a$ are univariate normal distributions.

Proposition 3.2. *Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$, and let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ be univariate normal distributions, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$. Then the distribution D is* ideal *for equal opportunity (see Definition 3.1) if and only if*

$$\frac{\mu_{01} - \mu_{11}}{\sigma_{11}} = \frac{\mu_{00} - \mu_{10}}{\sigma_{10}}, \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\sigma_{10}}{\sigma_{00}}, \quad \frac{q_{10}}{q_{00}} = \frac{q_{11}}{q_{01}}.$$

It is interesting to note that the same parametric conditions imply that the Bayes optimal classifier on the corresponding distribution simultaneously satisfies multiple fairness criteria, viz., demographic parity, equal opportunity, and equalized odds. Moreover, the same condition works for both univariate Gaussian and log-normal distributions. Using our proof technique, it is easy to derive similar conditions for other parametric families too.

Kamiran & Calders (2012) reweighing method essentially reweighs q_{ia} by a multiplicative factor of $\Pr(Y=i)\Pr(A=a)/\Pr(Y=i, A=a)$. Let us call the resulting probabilities $\tilde{q}_{ia}$. Using $\Pr(Y=i) = q_{i0} + q_{i1}$, $\Pr(A=a) = q_{0a} + q_{1a}$ and $\Pr(Y=i, A=a) = q_{ia}$, we get

$$\tilde{q}_{ia} \propto q_{ia} \frac{(q_{i0}+q_{i1})(q_{0a}+q_{1a})}{q_{ia}} \implies \frac{\tilde{q}_{10}}{\tilde{q}_{00}} = \frac{\tilde{q}_{11}}{\tilde{q}_{01}} = \frac{q_{10}+q_{11}}{q_{00}+q_{01}}.$$

It is the same condition on q_{ia}'s stated in Proposition 3.3. Thus, our result can be thought of as a second stage pre-processing of P_{ia} distributions after applying the reweighing of Kamiran & Calders (2012) to q_{ia}'s in the first stage. Now we state our result for multivariate Gaussians.

Proposition 3.3. *Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y=i, A=a)$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $X|Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions with mean $\mu_{ia} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ia} \in \mathbb{R}^{d\times d}$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. If $q_{10}/q_{00} = q_{11}/q_{01}$ and the means μ_{ia} and the covariance matrices Σ_{ia} satisfy*

$$\Sigma_{10}^{-1/2}(\mu_{10}-\mu_{00}) = \Sigma_{11}^{-1/2}(\mu_{11}-\mu_{01}) \quad \textit{and} \qquad \Sigma_{10}^{1/2}\Sigma_{00}^{-1}\Sigma_{10}^{1/2} = \Sigma_{11}^{1/2}\Sigma_{01}^{-1}\Sigma_{11}^{1/2}$$

then the group-aware Bayes optimal classifier on D satisfies equal opportunity.

Proposition 3.2 shows that our parametric condition is equivalent to $\Delta_C(h^*_C, D) = 0$, for all cost matrices $C \in R^{2\times 2}$. When we use a fixed cost matrix for 0-1 loss, and consider the Bayes optimal classifier in Proposition 3.3, our parametric condition is sufficient but not always necessary. However, the same condition ensures the Bayes optimal classifier to satisfy multiple fairness criteria simultaneously, viz., demographic parity, equal opportunity, equalized odds.

Remark 3.4. As an interesting consequence, our conditions on μ_{ia} and Σ_{ia} imply $D_{\mathrm{KL}}\left(\tilde{P}_{00}||\tilde{P}_{01}\right) = D_{\mathrm{KL}}\left(\tilde{P}_{10}||\tilde{P}_{11}\right)$. When the classes are balanced, the error rate of the Bayes optimal classifier on group $A=a$ in $\tilde{D}$ equals $\frac{1}{2}\left(1 - d_{\mathrm{TV}}(\tilde{P}_{0a}, \tilde{P}_{1a})\right)$, where d_{TV} denotes the total variation distance Nielsen (2014). Thus, achieving $d_{\mathrm{TV}}(\tilde{P}_{00}, \tilde{P}_{10}) = d_{\mathrm{TV}}(\tilde{P}_{01}, \tilde{P}_{11})$ ensures equal error rates across both the groups. However, there is no closed form expression for the total variation distance between two univariate Gaussians, and KL-divergence can be thought of as a proxy using Pinsker's inequality Canonne (2023). This is similar to information theoretic argument used by Dutta et al. (2020), which is why their optimization cannot guarantee outcome fairness.

4 Finding The Nearest Optimal Distribution

When a given distribution D is not ideal, then a natural question is to find its nearest distribution $\tilde{D}$ that is ideal. We formulate this problem as follows.

$$\underset{\tilde{D}\ :\ \tilde{D} \text{ is ideal}}{\text{minimize}}\ D_{\mathrm{KL}}\left(\tilde{D}||D\right).$$

In the above optimization problem, the KL-divergence objective is well-known and convex but the constraint of D' being ideal is extremely non-trivial to express. We show that when the group and class-conditioned distributions $X \mid Y=i, A=a$ come from certain well-known parametric families of distributions, this constraint can be equivalently expressed as a constraint on the distribution parameters. We now give a concrete formulation of the optimization problem described in Section using the constraints derived in Proposition 3.3.

4.1 Affirmative Action

We first focus on a class of interventions for which solving the optimization program is efficient. We define *Affirmative Action* as changing the underprivileged group to obtain the ideal distributions where fairness and accuracy are in accord.

Theorem 4.1. *Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{10}/q_{00} = q_{11}/q_{01}$. Let*

$X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions, with mean $\mu_{ia} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ia} \in \mathbb{R}^{d\times d}$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y = i, A = a$ to $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$. Then in the case of Affirmative action (changing only $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$), we can efficiently minimize $D_{\mathrm{KL}}\left(\tilde{D}||D\right)$ as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$ subject to the constraints in Proposition 3.3, so that the Bayes optimal classifier on the optimal $\tilde{D}$ is guaranteed to be EO-fair.

While we show that the optimization program is convex, obtaining a closed-form expression for the change in means and covariances is extremely cumbersome for the general case. However, we can show how the closed form expressions for $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ for univariate distributions (Corollary C.2). Another intervention we can follow is to change all the subgroups of the given distribution. However, a quick check through the proof of Theorem 4.1 shows that this will lead to a non-convex program. However, just like Corollary C.2, we can show a reasonable intervention for the univariate case, where we change all four subgroups and search over a non-convex function using line search over a fairly large grid size. We demonstrate this in Proposition C.3.

Finally, we also consider another intervention where we match the first moment of the underprivileged group with the privileged group, inspired by the commonly studied Calders-Verwer gap Calders & Verwer (2010); Kamishima et al. (2012); Chen et al. (2019). The resulting program is convex and we specify the closed-form expressions for $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ in Proposition C.4. We will leverage these interventions in Section 5 to study their ability to obtain ideal fair and optimal distributions that are also close to the given distribution.

4.2 Consequences for Fairness and Accuracy

Let $\tilde{h}$ be the Bayes optimal classifier on our ideal distribution $\tilde{D}$ and let h be the Bayes optimal classifier on D. The error rate and fairness guarantees of $\tilde{h}$ can be translated approximately from $\tilde{D}$ to D as follows.

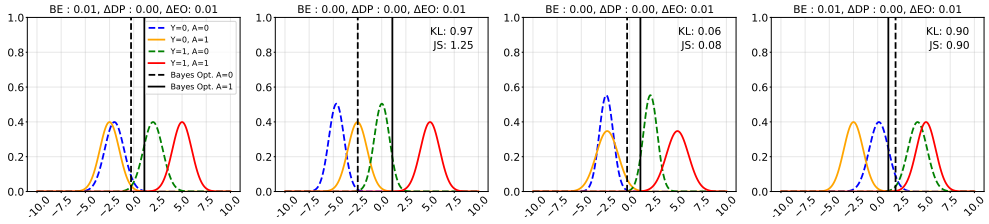
$$\begin{aligned}\text{err}(\tilde{h}, D) \leq \text{err}(\tilde{h}, \tilde{D}) + d_{\text{TV}}(\tilde{D}, D) &\leq \text{err}(h, D) + O(d_{TV}(\tilde{D}, D)) + d_{TV}(\tilde{D}, D) \\ &\leq \text{err}(h, D) + O(\sqrt{D_{\mathrm{KL}}\left(\tilde{D}||D\right)}.\end{aligned}$$

The first inequality follows from writing the error as expected 0-1 loss and using the definition of total variation distance. The second inequality follows from using a result of Kearns & Li (1993) on learning under malicious noise. The last line follows from Pinsker's inequality Canonne (2023). Thus, the optimal value of our optimization problem can be used to approximately translate the accuracy guarantee of $\tilde{h}$ from $\tilde{D}$ to D. A similar proof works for fairness.

5 Case Study on Gaussian Distributions

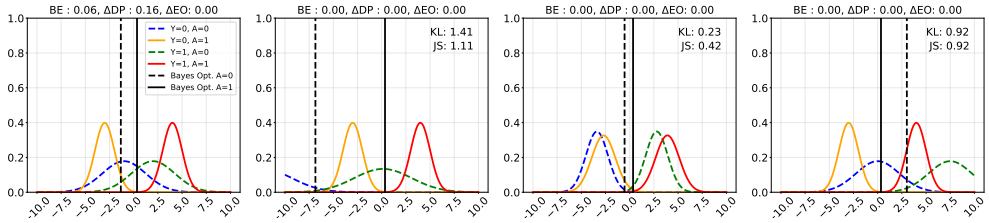
In this section, we modify a stylized setting of Gaussian distributions from previous work (see Definition 3.1 in Pierson et al. (2018), Section 5.3 in Bakalar et al. (2021)) to investigate the unfairness and the Bayes optimal error on the original and ideal distributions obtained through various interventions. Details of the setup can be found in Appendix D. The different interventions we try are as follows: (a) a simple Bayes optimal classifier on the original distribution without any correction, (b) Affirmative Action for Exact fairness (*EF-Affirmative*), where we change the underprivileged group (A=0) using the solution of the univariate KL divergence program from Corollary C.2, (c) Changing all subgroups for Exact fairness (*EF-All Subgroups*), where we change all subgroups to minimize the KL divergence with respect to the true distribution subject to exact fairness constraints from Proposition C.3, and (d) *Mean Matching*, where we minimize the KL divergence with respect to the true distribution subject to matching the means of the sensitive groups from Proposition C.4. Note that EF-All Subgroups lead to a non-convex optimization program, unlike the EF-Affirmative program. Therefore, we employ line search to approximate the factor γ that determines the optimal means and variances.

We measure the Demographic Parity (ΔDP) and Equal Opportunity (ΔEO) Difference of the Bayes optimal classifier on the new distributions, along with the KL and Jensen-Shannon (JS) divergence



(a) Original Distribution (b) EF-Affirmative (c) EF-All Subgroups (d) Mean Matching

Figure 2: Comparison of Different Interventions when the original distribution is already fair. In this case, EF-All ensures that it stays close to the true distribution, as no intervention as required, while others relatively deviate.



(a) Original Distribution (b) EF-Affirmative (c) EF-All Subgroups (d) Mean Matching

Figure 3: Comparison of Different Interventions when the ΔDP on the original distribution is high. In this case, EF-All manages to stay close to the true distribution and achieves perfect fairness and error rate, while others deviate significantly.

with respect to the true distribution. For the case of univariate gaussians, we precisely know the Bayes Optimal Classifier and the threshold (from Lemma B.1), and therefore, we use that to plot the group-aware decision thresholds and report the Bayes Error (BE). Furthermore, ΔDP and ΔEO can be computed analytically using differences of Cumulative Distribution functions of the standard gaussian.

First, we look at a case where the Bayes optimal classifier is already fair (ΔEO is close to 0 while ΔDP=0) in Figure 2. The expected solution here should be that any intervention must leave the distribution as it is. EF-Affirmative intervention keeps the unfairness and error rate numbers as it is, but deviates from the true distribution, as indicated by the KL/JS divergences. However, the EF-All intervention only makes major changes to variances and stays close to the true distribution. The Mean Matching intervention shifts both the under-privileged subgroups and strays away from the true distribution, as indicated by relatively high KL/JS values. We next construct a distribution where ΔDP is very high in Figure 3. Here, the affirmative action intervention transforms both the under-privileged subgroups to high-variance ones, which results in a reduction of BE and ΔDP, but at the cost of and high KL/JS-divergence with respect to the true distribution. However, the EF-All intervention simply tries to match the variances of under-privileged and privileged subgroups and, as a result, archives perfect fairness and accuracy while staying close to the true distribution. Mean Matching is very similar to EF-Affirmative in this case and, as a result, has relatively high KL/JS numbers. In Appendix D, we study the case of symmetric and shifted subgroup distribution and a different threshold in Figures 4 and 5 respectively.

6 Discussion and Future Work

In this work, we approach the problem of fair classification through the lens of *ideal* distributions. We first define what it means for a distribution to be ideal for fairness and Bayes optimal classification and then demonstrate that for well-known parametric families of distributions. We demonstrated when such an optimization problem is feasible and efficiently solvable. We show that these interventions can steer the given distribution to achieve perfect fairness and accuracy while staying close to the given distribution in many cases. Some important future directions include generalizing the above results for approximate fairness and studying the feasibility of the ideal distribution optimization program with finitely many samples or a bounded distance away from the given distribution.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. Fairness on the ground: Applying algorithmic fairness approaches to production systems. *CoRR*, abs/2103.06172, 2021. URL https://arxiv.org/abs/2103.06172.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in AI. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL https://aka.ms/fairlearn-whitepaper.

Sumon Biswas and Hridesh Rajan. Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline. In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2021, pp. 981–993, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450385626. doi: 10.1145/3468264.3468536. URL https://doi.org/10.1145/3468264.3468536.

Christina Hastings Blow, Lijun Qian, Camille Gibson, Pamela Obiomon, and Xishuang Dong. Comprehensive validation on reweighting samples for bias mitigation via aif360. *Applied Sciences*, 14 (9), 2024. ISSN 2076-3417. doi: 10.3390/app14093826. URL https://www.mdpi.com/2076-3417/14/9/3826.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? In *Symposium on Foundations of Responsible Computing (FORC)*, 2019.

Avrim Blum, Princewill Okoroafor, Aadirupa Saha, and Kevin Stangl. On the vulnerability of fairness constrained learning to malicious noise. *arXiv preprint arXiv:2307.11892*, 2023.

Stephen Boyd. Convex optimization. *Cambridge UP*, 2004.

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

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf.

Clément L. Canonne. A short note on an inequality between kl and tv, 2023. URL https://arxiv.org/abs/2202.07198.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning*, pp. 1349–1361. PMLR, 2021.

Mattia Cerrato, Marius Köppel, Philipp Wolf, and Stefan Kramer. 10 years of fair representations: Challenges and opportunities, 2024. URL https://arxiv.org/abs/2407.03834.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

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.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. Biased programmers? or biased data? a field experiment in operationalizing ai ethics. In *Proceedings of the 21st ACM Conference on Economics and Computation*, EC '20, pp. 679–681, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379755. doi: 10.1145/3391403.3399545. URL https://doi.org/10.1145/3391403.3399545.

Luc Devroye, László Györfi, and Gábor Lugosi. *The Bayes Error*, pp. 9–20. Springer, 1996. ISBN 978-1-4612-0711-5. doi: 10.1007/978-1-4612-0711-5_2.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pp. 2803–2813. PMLR, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, 2012. ISBN 9781450311151.

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

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. The (im)possibility of fairness: different value systems require different mechanisms for fair decision making. *Commun. ACM*, 64(4):136–143, March 2021. ISSN 0001-0782. doi: 10.1145/3433949. URL https://doi.org/10.1145/3433949.

Gene H Golub. Some modified matrix eigenvalue problems. *SIAM review*, 15(2):318–334, 1973.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, 2016. ISBN 9781510838819.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pp. 35–50. Springer, 2012.

Michael Kearns and Ming Li. Learning in the presence of malicious errors. *SIAM Journal on Computing*, 22(4):807–837, 1993. doi: 10.1137/0222052. URL https://doi.org/10.1137/0222052.

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.

Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. Impossibility results for fair representations, 2021. URL https://arxiv.org/abs/2107.03483.

Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, pp. 1088–1097, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539302. URL https://doi.org/10.1145/3534678.3539302.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/madras18a.html.

Sandra Gabriel Mayson. Bias in, bias out. *Yale Law Journal*, 2218, 2019. URL https://www.yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf.

Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pp. 263–270, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618.3317964. URL https://doi.org/10.1145/3306618.3317964.

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

Frank Nielsen. Generalized bhattacharyya and chernoff upper bounds on bayes error using quasi-arithmetic means. *Pattern Recognition Letters*, 42:25–34, 2014. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2014.01.002. URL https://www.sciencedirect.com/science/article/pii/S0167865514000166.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting discrimination. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 96–105. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/pierson18a.html.

Drago Plečko and Nicolai Meinshausen. Fair data adaptation with quantile preservation. *Journal of Machine Learning Research*, 21(242):1–44, 2020. URL http://jmlr.org/papers/v21/19-966.html.

Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. fairadapt: Causal reasoning for fair data preprocessing. *Journal of Statistical Software*, 110(4):1–35, 2024. doi: 10.18637/jss.v110.i04. URL https://www.jstatsoft.org/index.php/jss/article/view/v110i04.

Ashesh Rambachan and Jonathan Roth. Bias In, Bias Out? Evaluating the Folk Wisdom. In Aaron Roth (ed.), *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 6:1–6:15, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-142-9. doi: 10.4230/LIPIcs.FORC.2020.6. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.FORC.2020.6.

Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6(none):958 – 992, 2012. doi: 10.1214/12-EJS699. URL https://doi.org/10.1214/12-EJS699.

Mohit Sharma and Amit Jayant Deshpande. How far can fairness constraints help recover from biased data? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

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

Shashank Singh and Justin Khim. Optimal binary classification beyond accuracy. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=pm8Y8unXkkJ.

Zikai Xiong, Niccolò Dalmasso, Alan Mishler, Vamsi K. Potluru, Tucker Balch, and Manuela Veloso. Fairwasp: fast and optimal fair wasserstein pre-processing. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i14.29545. URL https://doi.org/10.1609/aaai.v38i14.29545.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/zemel13.html.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Fair bayes-optimal classifiers under predictive parity. *Advances in Neural Information Processing Systems*, 35:27692–27705, 2022.

A Related Work

A.1 Fair Pre-processing

Kamiran & Calders (2012) propose simple heuristics to pre-process data for binary fair classification with binary groups. Their most popular heuristic is to reweigh the data points in class i and group a by $\Pr(\text{class } i)\Pr(\text{group } a)/\Pr(\text{class } i \text{ and group } a)$. Note that this reweighing is independent of feature distribution, and hence cannot provide any provable guarantee on fairness when we maximize accuracy after reweighing. Calmon et al. (2017) formulate fair pre-precessing as an optimization problem to learn a data transformation that minimizes distance to the given distribution subject to bounds on discrimination (or group unfairness) and distortion control (or individual unfairness). They show conditions under which this optimization is convex and can be solved efficiently, however, it can be infeasible when group fairness (near equal outcomes for two groups) and individual fairness (similar outcomes for similar individuals across groups) cannot be satisfied simultaneously Friedler et al. (2021). Jiang & Nachum (2020) propose reweighing as a way to correct label bias in data. Plečko & Meinshausen (2020) and Plečko et al. (2024) propose a fair adaptation methods based on causal model of the data. More recently, Xiong et al. (2024) reformulate the task of reweighing given data for fair pre-processing as a large-scale mixed-integer program and propose an efficient algorithm to solve it via cutting-plane method. In general, pre-processing is practical and useful and comes as part of popular fairness toolkits along with in-processing and post-processing methods Bellamy et al. (2018); Bird et al. (2020). Even the simple reweighing of Kamiran & Calders (2012) that comes without any provable guarantees is surprisingly effective at bias mitigation on standard datasets in fair machine learning Sharma et al. (2023); Blow et al. (2024); Xiong et al. (2024).

A.2 Ideal Distributions and Fair Representation Learning

Dutta et al. (2020) characterize a similar objective using Chernoff Information (see Cover (1999)) and formulate an optimization program to find the nearest distribution in KL-divergence on which

the Chernoff Information gap between two group-conditional feature distributions vanishes. Their optimization problem is not known to be efficiently solvable and the fairness guarantees in terms of Chernoff Information gap does not translate easily to standard fairness metrics such as demographic parity, equal opportunity etc. In contrast, we formulate an optimization problem to find the nearest *ideal* distribution in KL-divergence to given distribution and give efficient algorithms to solve it for various parametric families of distributions.

For completeness, we want to also make the reader aware of a long line of work on fair representation learning where the data transformations can map the distributions to another space Zemel et al. (2013); Madras et al. (2018); McNamara et al. (2019); Liu et al. (2022); Cerrato et al. (2024). Our work is not directly related but can potentially be used to refine fair representations to achieve provable and exact fairness guarantees.

B Proofs for Section 3

We will require a helper result about threshold classifiers to prove our next set of results.

Lemma B.1. *Let* $\eta(x,a) = \Pr(Y=1|X=x,A=a)$, $q_{ia} = \Pr(Y=i,A=a)$ *and* $p_{ia}(x) = \Pr\left(X=x \;\middle|\; Y=i,A=a\right)$. *Then the Bayes optimal classifier can be written as* $h^*(x,a) = \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geq \log\frac{q_{0a}}{q_{1a}}\right)$.

Proof. Let $\eta(x,a) = \Pr(Y=1|X=x,A=a)$, $q_{ia} = \Pr(Y=i,A=a)$ and $p_{ia}(x) = \Pr\left(X=x \;\middle|\; Y=i,A=a\right)$. We consider group-aware threshold classifiers on D of the form $h_t(x,a) = \mathbb{I}(\eta(x,a) \geq t)$, which can be equivalently written as

$$\begin{aligned}
h_t(x,a) &= \mathbb{I}(\eta(x,a) \geq t) \\
&= \mathbb{I}\left(\Pr\left(Y=1 \;\middle|\; X=x,A=a\right) \geq t\right) \\
&= \mathbb{I}\left(\frac{\Pr\left(Y=1 \;\middle|\; X=x,A=a\right)}{\Pr\left(Y=0 \;\middle|\; X=x,A=a\right)} \geq \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{\Pr(Y=1,X=x,A=a)}{\Pr(Y=0,X=x,A=a)} \geq \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{\Pr\left(X=x \;\middle|\; Y=1,A=a\right)\Pr(Y=1,A=a)}{\Pr\left(X=x \;\middle|\; Y=0,A=a\right)\Pr(Y=0,A=a)} \geq \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{p_{1a}(x)}{p_{0a}(x)} \geq \frac{t}{1-t}\cdot\frac{q_{0a}}{q_{1a}}\right) \\
&= \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geq \log\frac{t}{1-t} + \log\frac{q_{0a}}{q_{1a}}\right).
\end{aligned}$$

It is well-known that the group-aware Bayes optimal classifier $h^* = h_{1/2}$ by setting $t = 1/2$, or equivalently,

$$h^*(x,a) = h_{1/2}(x,a) = \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geq \log\frac{q_{0a}}{q_{1a}}\right).$$

□

Proof. (Proof of Proposition 3.2) For any cost matrix $C \in \mathbb{R}^{2\times 2}$, the group-aware classifier that minimizes its corresponding cost-sensitive risk is given by $\mathbb{I}(\eta(x,a) \geq t_C)$, for a threshold $t_C = (c_{10}-c_{00})/(c_{10}-c_{00}+c_{01}-c_{11}) \in [0,1]$; see Equation (2) in Elkan (2001) and Scott (2012). The distribution D is *ideal* for equal opportunity if $\Pr\left(\eta(X,A) \geq t \;\middle|\; Y=i,A=0\right) = \Pr\left(\eta(X,A) \geq t \;\middle|\; Y=i,A=1\right)$, for all thresholds $t \in [0,1]$ and $i \in \{0,1\}$. Since the CDFs are identical, the random variables $\eta(X,A) \;\middle|\; Y=i,A=0$ and $\eta(X,A) \;\middle|\; Y=i,A=1$ must be

identical. Note that

$$
\begin{aligned}
\eta(x,a) &= \Pr\left(Y=1 \;\middle|\; X=x, A=a\right) \\
&= \frac{\Pr\left(Y=1, X=x, A=a\right)}{\sum_{i=0}^{1}\Pr\left(Y=i, X=x, A=a\right)} \\
&= \frac{\Pr\left(Y=1, A=a\right)\Pr\left(X=x \;\middle|\; Y=1, A=a\right)}{\sum_{i=0}^{1}\Pr\left(Y=i, A=a\right)\Pr\left(X=x \;\middle|\; Y=i, A=a\right)} \\
&= \frac{q_{1a}P_{1a}(x)}{\sum_{i=0}^{1} q_{ia}P_{ia}(x)} \\
&= \frac{q_{1a}\sigma_{1a}^{-1}\exp\left(-\frac{(x-\mu_{1a})^2}{2\sigma_{1a}^2}\right)}{\sum_{i=0}^{1} q_{ia}\sigma_{ia}^{-1}\exp\left(-\frac{(x-\mu_{ia})^2}{2\sigma_{ia}^2}\right)} \\
&= \frac{1}{1+\exp\left(\frac{(x-\mu_{1a})^2}{2\sigma_{1a}^2} - \frac{(x-\mu_{0a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)} \\
&= \frac{1}{1+\exp\left(\frac{(\mu_{ia}+r\sigma_{ia}-\mu_{1a})^2}{2\sigma_{1a}^2} - \frac{(\mu_{ia}+r\sigma_{ia}-\mu_{0a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)} \\
&= \begin{cases}
\dfrac{1}{1+\exp\left(\frac{1}{2}\left(\frac{\sigma_{0a}^2}{\sigma_{1a}^2}-1\right)r^2 - \frac{\sigma_{0a}(\mu_{1a}-\mu_{0a})}{\sigma_{1a}^2}r + \frac{(\mu_{1a}-\mu_{0a})^2}{2\sigma_{1a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)}, & \text{for } i=0 \\
\dfrac{1}{1+\exp\left(\frac{1}{2}\left(1-\frac{\sigma_{1a}^2}{\sigma_{0a}^2}\right)r^2 - \frac{\sigma_{1a}(\mu_{0a}-\mu_{1a})}{\sigma_{0a}^2}r + \frac{(\mu_{0a}-\mu_{1a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)}, & \text{for } i=1.
\end{cases}
\end{aligned}
$$

If $X|Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$, then $X \;\middle|\; Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$. Thus, for $\eta(X,A) \;\middle|\; Y=i, A=0$ and $\eta(X,A) \;\middle|\; Y=i, A=1$ to be identical, we must have

$$
\begin{aligned}
&\frac{1}{2}\left(\frac{\sigma_{00}^2}{\sigma_{10}^2}-1\right)R^2 - \frac{\sigma_{00}(\mu_{10}-\mu_{00})}{\sigma_{10}^2}R + \frac{(\mu_{10}-\mu_{00})^2}{2\sigma_{10}^2} + \log\frac{q_{00}\sigma_{10}}{q_{10}\sigma_{00}} \quad \text{and} \\
&\frac{1}{2}\left(\frac{\sigma_{01}^2}{\sigma_{11}^2}-1\right)R^2 - \frac{\sigma_{01}(\mu_{11}-\mu_{01})}{\sigma_{11}^2}R + \frac{(\mu_{11}-\mu_{01})^2}{2\sigma_{11}^2} + \log\frac{q_{01}\sigma_{11}}{q_{11}\sigma_{01}}
\end{aligned}
$$

as identically distributed for $R \sim \mathcal{N}(0,1)$. Similarly, we must also have

$$
\begin{aligned}
&\frac{1}{2}\left(1-\frac{\sigma_{10}^2}{\sigma_{00}^2}\right)R^2 - \frac{\sigma_{10}(\mu_{00}-\mu_{10})}{\sigma_{00}^2}R + \frac{(\mu_{00}-\mu_{10})^2}{2\sigma_{00}^2} + \log\frac{q_{00}\sigma_{10}}{q_{10}\sigma_{00}} \quad \text{and} \\
&\frac{1}{2}\left(1-\frac{\sigma_{11}^2}{\sigma_{01}^2}\right)R^2 - \frac{\sigma_{11}(\mu_{01}-\mu_{11})}{\sigma_{01}^2}R + \frac{(\mu_{01}-\mu_{11})^2}{2\sigma_{01}^2} + \log\frac{q_{01}\sigma_{11}}{q_{11}\sigma_{01}}
\end{aligned}
$$

as identically distributed for $R \sim \mathcal{N}(0,1)$. Therefore, we must have

$$
\frac{\mu_{01}-\mu_{11}}{\sigma_{11}} = \frac{\mu_{00}-\mu_{10}}{\sigma_{10}} \quad \text{and} \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\sigma_{10}}{\sigma_{00}} \quad \text{and} \quad \frac{q_{10}}{q_{00}} = \frac{q_{11}}{q_{01}}.
$$

In the other direction, it is easier to prove that the above conditions imply the distribution to be ideal. It can be proved by simply backtracking the steps above. □

Proof. **(Proof of Proposition 3.3)** From Lemma B.1, the group-aware Bayes optimal classifier can be written as

$$
h^*(x,a) = \mathbb{I}\left(\eta(x,a) \geq \frac{1}{2}\right) = \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geq \log\frac{q_{0a}}{q_{1a}}\right).
$$

The EO-fairness condition $\Pr\left(h^*(X,A)=1 \mid Y=1, A=0\right) = \Pr\left(h^*(X,A)=1 \mid Y=1, A=1\right)$ means

$$\Pr\left(\log\frac{p_{10}(X)}{p_{00}(X)} \geq \log\frac{q_{00}}{q_{10}} \mid Y=1, A=0\right) = \Pr\left(\log\frac{p_{11}(X)}{p_{01}(X)} \geq \log\frac{q_{01}}{q_{11}} \mid Y=1, A=0\right).$$

Since $X|Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ are multivariate Normal distributions for $i, a \in \{0,1\}$, their probability densities are

$$p_{ia}(x) = (2\pi)^{-d/2}\det(\Sigma_{ia})^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_{ia})^T\Sigma_{ia}^{-1}(x-\mu_{ia})\right).$$

Now we can write

$$\begin{aligned}
\log\frac{p_{1a}(x)}{p_{0a}(x)} &= \frac{1}{2}\left((x-\mu_{0a})^T\Sigma_{0a}^{-1}(x-\mu_{0a}) - (x-\mu_{1a})^T\Sigma_{1a}^{-1}(x-\mu_{1a}) + \log\det(\Sigma_{0a}) - \log\det(\Sigma_{1a})\right) \\
&= \frac{1}{2}\left((\Sigma_{1a}^{1/2}r+\mu_{1a}-\mu_{0a})^T\Sigma_{0a}^{-1}(\Sigma_{1a}^{1/2}r+\mu_{1a}-\mu_{0a}) - r^Tr - \log\det(\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2})\right) \\
&\qquad\qquad \text{by substituting } x = \Sigma_{1a}^{1/2}r+\mu_{1a}, \text{ where } r\sim\mathcal{N}(0, I_{d\times d}) \\
&= \frac{1}{2}\,r^T\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}r + (\mu_{1a}-\mu_{0a})^T\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}r - \frac{1}{2}\log\det(\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}) \\
&= \frac{1}{2}\,r^T\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}r + (\mu_{1a}-\mu_{0a})^T\Sigma_{1a}^{-1/2}\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}r - \frac{1}{2}\log\det(\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}).
\end{aligned}$$

Hence,

$$\begin{aligned}
&\Pr\left(\log\frac{p_{1a}(X)}{p_{0a}(X)} \geq \log\frac{q_{0a}}{q_{1a}} \mid Y=1, A=a\right) \\
&= \Pr\left(\frac{1}{2}\,R^T\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}R + (\mu_{1a}-\mu_{0a})^T\Sigma_{1a}^{-1/2}\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}R - \frac{1}{2}\log\det(\Sigma_{1a}^{1/2}\Sigma_{0a}^{-1}\Sigma_{1a}^{1/2}) \geq \log\frac{q_{0a}}{q_{1a}}\right),
\end{aligned}$$

for $R\sim\mathcal{N}(\bar{0}, I_{d\times d})$. Now if we have $\dfrac{q_{10}}{q_{00}} = \dfrac{q_{11}}{q_{01}}$ and

$$\Sigma_{10}^{-1/2}(\mu_{10}-\mu_{00}) = \Sigma_{11}^{-1/2}(\mu_{11}-\mu_{01}) \quad \text{and} \quad \Sigma_{10}^{1/2}\Sigma_{00}^{-1}\Sigma_{10}^{1/2} = \Sigma_{11}^{1/2}\Sigma_{01}^{-1}\Sigma_{11}^{1/2},$$

then the probability of the above event written in terms $R\sim\mathcal{N}(\bar{0}, I_{d\times d})$ becomes identical for $a\in\{0,1\}$. Hence, the Bayes optimal classifier satisfies equal opportunity. □

C Proofs for Section 4

We first derive the KL divergence between two distributions, where each subgroup in the distribution follows a multivariate normal distribution.

Lemma C.1. *Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y=i, A=a)$, for $i\in\{0,1\}$ and $a\in\{0,1\}$. Let $X|Y=i, A=a\sim\mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions with mean $\mu_{ia}\in\mathbb{R}^d$ and covariance matrix $\Sigma_{ia}\in\mathbb{R}^{d\times d}$, for $i\in\{0,1\}$ and $a\in\{0,1\}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y=i, A=a$ to $\tilde{X}|Y=i, A=a\sim\mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$. Then,*

$$\begin{aligned}
D_{\mathrm{KL}}\left(\tilde{D}||D\right) = -\frac{d}{2} &+ \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) \\
&+ \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) - \log\det(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia})\right).
\end{aligned}$$

Proof.

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(\tilde{D}||D\right) \\
&= \sum_{(x,i,a)} \Pr\left(\tilde{X}=x,\tilde{Y}=i,\tilde{A}=a\right) \log \frac{\Pr\left(\tilde{X}=x,\tilde{Y}=i,\tilde{A}=a\right)}{\Pr\left(X=x,Y=i,A=a\right)} \\
&= \sum_{(x,i,a)} \Pr\left(\tilde{Y}=i,\tilde{A}=a\right) \Pr\left(\tilde{X}=x \;\middle|\; \tilde{Y}=i,\tilde{A}=a\right) \log \frac{\Pr\left(\tilde{Y}=i,\tilde{A}=a\right) \Pr\left(\tilde{X}=x \;\middle|\; \tilde{Y}=i,\tilde{A}=a\right)}{\Pr\left(Y=y,A=a\right) \Pr\left(X=x \;\middle|\; Y=i,A=a\right)} \\
&= \sum_{(x,i,a)} \Pr\left(Y=i,A=a\right) \Pr\left(\tilde{X}=x \;\middle|\; Y=i,A=a\right) \log \frac{\Pr\left(Y=i,A=a\right) \Pr\left(\tilde{X}=x \;\middle|\; Y=i,A=a\right)}{\Pr\left(Y=i,A=a\right) \Pr\left(X=x \;\middle|\; Y=i,A=a\right)} \\
&= \sum_{(i,a)} q_{ia} \sum_{x} \Pr\left(\tilde{X}=x \;\middle|\; Y=i,A=a\right) \log \frac{\Pr\left(\tilde{X}=x \;\middle|\; Y=i,A=a\right)}{\Pr\left(X=x \;\middle|\; Y=i,A=a\right)} \\
&= \sum_{(i,a)} q_{ia} D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right)
\end{aligned}
$$

P_{ia} denotes the distribution of $X \;\middle|\; Y=i,A=a \sim \mathcal{N}(\mu_{ia},\Sigma_{ia})$ and $\tilde{P}_{ia}$ denotes the distribution of $\tilde{X} \;\middle|\; Y=i,A=a \sim \mathcal{N}(\tilde{\mu}_{ia},\tilde{\Sigma}_{ia})$. Their probability densities are

$$
\begin{aligned}
p_{ia}(x) &= (2\pi)^{-d/2} \det(\Sigma_{ia})^{-1/2} \exp\left(-\frac{1}{2}(x-\mu_{ia})^T \Sigma_{ia}^{-1}(x-\mu_{ia})\right) \quad \text{and} \\
\tilde{p}_{ia}(x) &= (2\pi)^{-d/2} \det(\tilde{\Sigma}_{ia})^{-1/2} \exp\left(-\frac{1}{2}(x-\tilde{\mu}_{ia})^T \tilde{\Sigma}_{ia}^{-1}(x-\tilde{\mu}_{ia})\right),
\end{aligned}
$$

respectively. Hence, the Kullback-Leibler divergence between $\tilde{P}_{ia}$ and P_{ia} can be written as

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right) \\
&= \mathbb{E}\log \frac{\tilde{p}_{ia}(\tilde{X})}{p_{ia}(\tilde{X})} \;\middle|\; Y=i,A=a \\
&= \frac{1}{2}\,\mathbb{E}(\tilde{X}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\mu_{ia}) - (\tilde{X}-\tilde{\mu}_{ia})^T\tilde{\Sigma}_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia}) - \log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}) \;\middle|\; Y=i,A=a \\
&= \frac{1}{2}\,\mathbb{E}(\tilde{X}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\mu_{ia}) \;\middle|\; Y=i,A=a - \frac{d}{2} - \frac{1}{2}\log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}) \\
&\qquad\qquad \text{using } \mathbb{E}(\tilde{X}-\tilde{\mu}_{ia})^T\tilde{\Sigma}_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia}) \;\middle|\; Y=i,A=a = \tilde{\Sigma}_{ia}^{-1} \bullet \tilde{\Sigma}_{ia} = \mathrm{tr}\left(I_{d\times d}\right) = d \\
&= \frac{1}{2}\,\mathbb{E}(\tilde{X}-\tilde{\mu}_{ia}+\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia}+\tilde{\mu}_{ia}-\mu_{ia}) \;\middle|\; Y=i,A=a - \frac{d}{2} + \frac{1}{2}\log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}) \\
&= \frac{1}{2}\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) + \frac{1}{2}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) - \frac{d}{2} + \frac{1}{2}\log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}).
\end{aligned}
$$

The Kullback-Leibler divergence between $\tilde{D}$ and D can now be written as

$$\begin{aligned}
D_{\mathrm{KL}}\left(\tilde{D}||D\right) &= \sum_{(i,a)} q_{ia} D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right) \\
&= \sum_{(i,a)} q_{ia}\left(\frac{1}{2}\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) + \frac{1}{2}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) - \frac{d}{2} + \frac{1}{2}\log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2})\right) \\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) + (\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \log\det(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2})\right) \\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) + \log\det(\Sigma_{ia}\tilde{\Sigma}_{ia}^{-1})\right) \\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) - \log\det(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia})\right).
\end{aligned}$$

□

Proof. (Proof for Theorem 4.1) Using Lemma C.1 and Proposition 3.3, our objective is to minimize

$$D_{\mathrm{KL}}\left(\tilde{D}||D\right) = -\frac{d}{2}+\frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia})+\frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right) - \log\det(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia})\right),$$

subject to the constraints

$$\tilde{\Sigma}_{10}^{-1/2}(\tilde{\mu}_{10}-\tilde{\mu}_{00}) = \tilde{\Sigma}_{11}^{-1/2}(\tilde{\mu}_{11}-\tilde{\mu}_{01}) \quad \text{and} \quad \tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{00}^{-1}\tilde{\Sigma}_{10}^{1/2} = \tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{01}^{-1}\tilde{\Sigma}_{11}^{1/2}.$$

Suppose $\tilde{\Sigma}_{i0}$ and $\tilde{\Sigma}_{i1}$ do not commute. The constraints can be equivalently rewritten as follows.

$$\tilde{\mu}_{10}-\tilde{\mu}_{00} = \tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{11}^{-1/2}(\tilde{\mu}_{11}-\tilde{\mu}_{01}) \quad \text{and} \quad \tilde{\Sigma}_{11}^{-1/2}\tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{00}^{-1}\tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{11}^{-1/2} = \tilde{\Sigma}_{01}^{-1}.$$

Let $\Gamma = \tilde{\Sigma}_{i0}^{1/2}\tilde{\Sigma}_{i1}^{-1/2}$. For any fixed positive semidefinite matrix $\Gamma \in \mathbb{R}^{d\times d}$, our optimization problem can be divided into two separate parts that minimize

$$\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) \quad \text{subject to} \quad \tilde{\mu}_{10}-\tilde{\mu}_{00} = \Gamma(\tilde{\mu}_{11}-\tilde{\mu}_{01})$$

over $\tilde{\mu}_{ia} \in \mathbb{R}^d$, for $i, a \in \{0,1\}$, and minimize (after substituting $\tilde{\Sigma}_{i0}^{1/2} = \Gamma\tilde{\Sigma}_{i1}^{1/2}$)

$$\sum_{i=0}^{1} q_{i0}\left(\mathrm{tr}\left(\Sigma_{i0}^{-1}\left(\Gamma\tilde{\Sigma}_{i0}^{1/2}\right)^2\right) - \log\det(\Sigma_{i0}^{-1}\left(\Gamma\tilde{\Sigma}_{i0}^{1/2}\right)^2\right) + q_{i1}\left(\mathrm{tr}\left(\Sigma_{i1}^{-1}\tilde{\Sigma}_{i1}\right) - \log\det(\Sigma_{i1}^{-1}\tilde{\Sigma}_{i1})\right),$$

subject to $\Gamma\tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{00}^{-1}\Gamma = \tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{01}^{-1}$

over symmetric, positive semidefinite matrix-valued variable $\tilde{\Sigma}_{i1} \in \mathbb{R}^{d\times d}$, for $i \in \{0,1\}$. The first optimization in $\tilde{\mu}_{ia}$ is a constrained eigenvalue problem with linear constraints, i.e., minimize $x^TAx + x^Tb$ subject to $x^Tc = e$ Golub (1973).

Let's consider the case of *Affirmative Action*, where we only change the means $\tilde{\mu}_{i0}$ and the covariance matrices $\tilde{\Sigma}_{i0}$ for the underprivileged group but keep those for the privileged group unchanged, i.e., $\tilde{\mu}_{i1} = \mu_{i1}$ and $\tilde{\Sigma}_{i1} = \Sigma_{i1}$. In that case, $\tilde{\Sigma}_{00}^{1/2} = \Gamma\Sigma_{01}^{1/2}$ and $\tilde{\Sigma}_{10}^{1/2} = \Gamma\Sigma_{11}^{1/2}$ get fixed. By substituting $\tilde{\mu}_{10} = \tilde{\mu}_{00} + \Gamma(\tilde{\mu}_{11}-\tilde{\mu}_{01}) = \tilde{\mu}_{00} + \Gamma(\mu_{11}-\mu_{01})$, we only need to optimize

$$q_{00}(\tilde{\mu}_{00}-\mu_{00})^T\Sigma_{00}^{-1}(\tilde{\mu}_{00}-\mu_{00})+q_{10}(\tilde{\mu}_{00}+\Gamma(\mu_{11}-\mu_{01})-\mu_{10})^T\Sigma_{10}^{-1}(\tilde{\mu}_{00}+\Gamma(\mu_{11}-\mu_{01})-\mu_{10}),$$

or equivalently (ignoring the terms independent of $\tilde{\mu}_{00}$),

$$\tilde{\mu}_{00}^T\left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)\tilde{\mu}_{00} - 2\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11}-\mu_{01})\right)^T\tilde{\mu}_{00}.$$

This is a convex objective in $\tilde{\mu}_{00}$ because its Hessian is positive semidefinite, i.e., $q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1} \succcurlyeq 0$ Boyd (2004). By equating the gradient to zero, we get the optimal solution for $\tilde{\mu}_{00}$,

and we denote it by $\mu^*_{00}(\Gamma)$. Thus, the optimal solutions $\mu^*_{00}(\Gamma), \mu^*_{10}(\Gamma), \Sigma^*_{00}(\Gamma), \Sigma^*_{10}(\Gamma)$ for a fixed positive semidefinite $\Gamma \in \mathbb{R}^{d\times d}$ are given by

$$\mu^*_{00}(\Gamma) = \left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)^{-1}\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11}-\mu_{01})\right) \quad \text{and}$$
$$\mu^*_{10}(\Gamma) = \left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)^{-1}\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11}-\mu_{01})\right) + \Gamma(\mu_{11}-\mu_{01})$$
$$\Sigma^*_{00}(\Gamma) = (\Gamma\Sigma_{01}^{1/2})^2$$
$$\Sigma^*_{10}(\Gamma) = (\Gamma\Sigma_{11}^{1/2})^2.$$

By substituting these, when we look at the objective as a function of a positive semidefinite matrix-valued variable Γ, it turns out to be convex. This requires rewriting the expressions using the identities $\operatorname{tr}(AB) = \operatorname{tr}(BA), \det(AB) = \det(A)\det(B)$, and most importantly, $\operatorname{tr}(AXBX) = \operatorname{tr}\left((A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T\right)$ and $\log\det(AXBX) = \log\det\left(A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T\right)$, for symmetric, positive semidefinite matrices A, B, X Petersen et al. (2008). The convexity of the objective in Γ follows from the convexity of $\operatorname{tr}(AXBX)$ and $-\log\det(X)$ for matrix-valued variable X. Finally, we can solve it efficiently to get the optimal Γ^*. □

Corollary C.2. *(Affirmative Action for the case of univariate distributions) For the case where $X|Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ are univariate normal distributions, for $i \in \{0,1\}$ and $a \in \{0,1\}$, the optimal distribution $\tilde{D}$ from Theorem 4.1 can be written down as:*

$$\tilde{\sigma}_{i0} = \gamma^*\sigma_{i1}, \quad \tilde{\mu}_{00} = \tilde{\mu}_{10} + \gamma^*(\mu_{01}-\mu_{11}), \text{ and}$$
$$\tilde{\mu}_{10} = \frac{\left(q_{00}\dfrac{\mu_{00}-\gamma^*(\mu_{01}-\mu_{11})}{\sigma^2_{00}} + q_{10}\dfrac{\mu_{10}}{\sigma^2_{10}}\right)}{\left(\dfrac{q_{00}}{\sigma^2_{00}} + \dfrac{q_{10}}{\sigma^2_{10}}\right)},$$

where γ^* is a function of the original distribution parameters only.

Proof. (Proof of Corollary C.2)

$$D_{\mathrm{KL}}\left(\tilde{D}||D\right)$$
$$= \sum_{(x,i,a)} \Pr\left(\tilde{X}=x, \tilde{Y}=i, \tilde{A}=a\right)\log\frac{\Pr\left(\tilde{X}=x, \tilde{Y}=i, \tilde{A}=a\right)}{\Pr\left(X=x, Y=i, A=a\right)}$$
$$= \sum_{(x,i,a)} \Pr\left(\tilde{Y}=i, \tilde{A}=a\right)\Pr\left(\tilde{X}=x \;\middle|\; \tilde{Y}=i, \tilde{A}=a\right)\log\frac{\Pr\left(\tilde{Y}=i, \tilde{A}=a\right)\Pr\left(\tilde{X}=x \;\middle|\; \tilde{Y}=i, \tilde{A}=a\right)}{\Pr\left(Y=y, A=a\right)\Pr\left(X=x \;\middle|\; Y=i, A=a\right)}$$
$$= \sum_{(x,i,a)} \Pr\left(Y=i, A=a\right)\Pr\left(\tilde{X}=x \;\middle|\; Y=i, A=a\right)\log\frac{\Pr\left(Y=i, A=a\right)\Pr\left(\tilde{X}=x \;\middle|\; Y=i, A=a\right)}{\Pr\left(Y=i, A=a\right)\Pr\left(X=x \;\middle|\; Y=i, A=a\right)}$$
$$= \sum_{(i,a)} q_{ia}\sum_x \Pr\left(\tilde{X}=x \;\middle|\; Y=i, A=a\right)\log\frac{\Pr\left(\tilde{X}=x \;\middle|\; Y=i, A=a\right)}{\Pr\left(X=x \;\middle|\; Y=i, A=a\right)}$$
$$= \sum_{(i,a)} q_{ia} D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right)$$

P_{ia} denotes the distribution of $X \;\middle|\; Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ and $\tilde{P}_{ia}$ denotes the distribution of $\tilde{X} \;\middle|\; Y=i, A=a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\sigma}^2_{ia})$. Their probability densities are

$$p_{ia}(x) = \frac{1}{x\sigma_{ia}\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_{ia})^2}{2\sigma^2_{ia}}\right) \quad \text{and} \quad \tilde{p}_{ia}(x) = \frac{1}{x\tilde{\sigma}_{ia}\sqrt{2\pi}}\exp\left(-\frac{(x-\tilde{\mu}_{ia})^2}{2\tilde{\sigma}^2_{ia}}\right),$$

respectively. Hence,

$$
\begin{aligned}
D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right) &= \mathrm{E}\left[\log\frac{\tilde{p}_{ia}(\tilde{X})}{p_{ia}(\tilde{X})} \mid Y=i, A=a\right] \\
&= \mathrm{E}\left[\frac{(\tilde{X}-\mu_{ia})^2}{2\sigma_{ia}^2} - \frac{(\tilde{X}-\tilde{\mu}_{ia})^2}{2\tilde{\sigma}_{ia}^2} + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \mid Y=i, A=a\right] \\
&= \mathrm{E}\left[\left(\frac{1}{2\sigma_{ia}^2} - \frac{1}{2\tilde{\sigma}_{ia}^2}\right)\tilde{X}^2 + \left(\frac{\tilde{\mu}_{ia}}{\tilde{\sigma}_{ia}^2} - \frac{\mu_{ia}}{\sigma_{ia}^2}\right)\tilde{X} + \left(\frac{\mu_{ia}^2}{2\sigma_{ia}^2} - \frac{\tilde{\mu}_{ia}^2}{2\tilde{\sigma}_{ia}^2}\right) + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \mid Y=i, A=a\right] \\
&= \left(\frac{1}{2\sigma_{ia}^2} - \frac{1}{2\tilde{\sigma}_{ia}^2}\right)(\tilde{\mu}_{ia}^2+\tilde{\sigma}_{ia}^2) + \left(\frac{\tilde{\mu}_{ia}}{\tilde{\sigma}_{ia}^2} - \frac{\mu_{ia}}{\sigma_{ia}^2}\right)\tilde{\mu}_{ia} + \left(\frac{\mu_{ia}^2}{2\sigma_{ia}^2} - \frac{\tilde{\mu}_{ia}^2}{2\tilde{\sigma}_{ia}^2}\right) + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \\
&= \frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2} + \frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2} + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}},
\end{aligned}
$$

using $\mathrm{E}\left[\log\tilde{X} \mid Y=i, A=a\right] = \tilde{\mu}_{ia}$ and $\mathrm{E}\left[(\log\tilde{X})^2 \mid Y=i, A=a\right] = \tilde{\mu}_{ia}^2+\tilde{\sigma}_{ia}^2$. Since we only change group $A=0$, we want to minimize

$$
\begin{aligned}
D_{\mathrm{KL}}\left(\tilde{D}||D\right) &= \sum_{i=0}^{1} q_{i0} D_{\mathrm{KL}}\left(\tilde{P}_{i0}||P_{i0}\right) \\
&= \sum_{i=0}^{1} q_{i0}\left(\frac{(\tilde{\mu}_{i0}-\mu_{i0})^2}{2\sigma_{i0}^2} + \frac{\tilde{\sigma}_{i0}^2-\sigma_{i0}^2}{2\sigma_{i0}^2} + \log\frac{\sigma_{i0}}{\tilde{\sigma}_{i0}}\right)
\end{aligned}
$$

as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ subject to the constraints

$$
\frac{\mu_{01}-\mu_{11}}{\sigma_{11}} = \frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\tilde{\sigma}_{10}} \quad \text{and} \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\tilde{\sigma}_{10}}{\tilde{\sigma}_{00}} \quad \text{and} \quad \tilde{\sigma}_{ia} \geq 0, \text{ for all } (i,a).
$$

Let's fix $\gamma \in \mathbb{R}_{\geq 0}$ and minimize

$$
\mathcal{L}_\gamma = \sum_{i=0}^{1} q_{i0}\left(\frac{(\tilde{\mu}_{i0}-\mu_{i0})^2}{2\sigma_{i0}^2} + \frac{\tilde{\sigma}_{i0}^2-\sigma_{i0}^2}{2\sigma_{i0}^2} + \log\frac{\sigma_{i0}}{\tilde{\sigma}_{i0}}\right)
$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the following constraints

$$
\frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\mu_{01}-\mu_{11}} = \frac{\tilde{\sigma}_{10}}{\sigma_{11}} = \frac{\tilde{\sigma}_{00}}{\sigma_{01}} = \gamma \quad \text{and} \quad \tilde{\sigma}_{ia} \geq 0, \text{ for all } (i,a).
$$

The objective $\mathcal{L}_\gamma$ is convex and for a fixed $\gamma \in \mathbb{R}_{\geq 0}$, the constraints on are linear in $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$. Let's denote the optimal solution for a fixed $\gamma \in \bar{\mathbb{R}}_{\geq 0}$ by $\mu_{i0}^*(\gamma)$ and $\sigma_{i0}^*(\gamma)$, for $i \in \{0,1\}$. For a fixed $\gamma \in \mathbb{R}_{\geq 0}$, the above constraints fix $\sigma_{i0}^*(\gamma) = \gamma\sigma_{i1}$, for $i \in \{0,1\}$, and by plugging in $\tilde{\mu}_{00} = \tilde{\mu}_{10} + \gamma(\mu_{01}-\mu_{11})$, we only need to minimize the following convex, quadratic objective in a single variable $\tilde{\mu}_{10}$,

$$
\text{minimize} \quad q_{00}\frac{(\tilde{\mu}_{10}+\gamma(\mu_{01}-\mu_{11})-\mu_{00})^2}{2\sigma_{00}^2} + q_{10}\frac{(\tilde{\mu}_{10}-\mu_{10})^2}{2\sigma_{10}^2}.
$$

By equating the derivative to zero, we get the optimal solution as

$$
\mu_{10}^*(\gamma) = \left(\frac{q_{00}}{\sigma_{00}^2} + \frac{q_{10}}{\sigma_{10}^2}\right)^{-1}\left(q_{00}\frac{\mu_{00}-\gamma(\mu_{01}-\mu_{11})}{\sigma_{00}^2} + q_{10}\frac{\mu_{10}}{\sigma_{10}^2}\right),
$$

and the optimal value at $\mu_{10}^*(\gamma)$ is (The min of ax^2+bx+c occurs at $x = \dfrac{-b}{2a}$ and has value $c - \dfrac{b^2}{4a}$)

$$
\begin{aligned}
&q_{00}\frac{(\gamma(\mu_{01}-\mu_{11})-\mu_{00})^2}{2\sigma_{00}^2} + q_{10}\frac{\mu_{10}^2}{2\sigma_{10}^2} - \frac{1}{2}\left(\frac{q_{00}}{\sigma_{00}^2} + \frac{q_{10}}{\sigma_{10}^2}\right)^{-1}\left(q_{00}\frac{\mu_{00}-\gamma(\mu_{01}-\mu_{11})}{\sigma_{00}^2} + q_{10}\frac{\mu_{10}}{\sigma_{10}^2}\right)^2 \\
&= \frac{1}{2}\left(\frac{q_{00}}{\sigma_{00}^2} + \frac{q_{10}}{\sigma_{10}^2}\right)^{-1}\frac{q_{00}q_{10}}{\sigma_{00}^2\sigma_{10}^2}\left((\mu_{00}-\mu_{10})-\gamma(\mu_{01}-\mu_{11})\right)^2 \\
&= \frac{1}{2}\left(\frac{\sigma_{00}^2}{q_{00}} + \frac{\sigma_{10}^2}{q_{10}}\right)^{-1}\left((\mu_{00}-\mu_{10})-\gamma(\mu_{01}-\mu_{11})\right)^2.
\end{aligned}
$$

By plugging in the optimal solution, the minimum value of $\mathcal{L}_\gamma$ for a fixed $\gamma \in \mathbb{R}_{\geq 0}$ is given by

$$\begin{aligned}
\mathcal{L}^*_\gamma &= \sum_{i=0}^{1} q_{i0} \left(\frac{(\mu^*_{i0}(\gamma) - \mu_{i0})^2}{2\sigma^2_{i0}} + \frac{\sigma^{*}_{i0}(\gamma)^2 - \sigma^2_{i0}}{2\sigma^2_{i0}} + \log \frac{\sigma_{i0}}{\sigma^*_{i0}(\gamma)} \right) \\
&= \frac{1}{2} \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)^{-1} ((\mu_{00} - \mu_{10}) - \gamma(\mu_{01} - \mu_{11}))^2 + q_{00} \frac{\gamma^2 \sigma^2_{01} - \sigma^2_{00}}{2\sigma^2_{00}} \\
&\qquad + q_{10} \frac{\gamma^2 \sigma^2_{11} - \sigma^2_{10}}{2\sigma^2_{10}} + (q_{00} + q_{10}) \log \frac{1}{\gamma} + q_{00} \log \frac{\sigma_{00}}{\sigma_{01}} + q_{10} \log \frac{\sigma_{10}}{\sigma_{11}} \\
&= \frac{1}{2} \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)^{-1} ((\mu_{00} - \mu_{10}) - \gamma(\mu_{01} - \mu_{11}))^2 + \frac{q_{00}}{2} \left(\gamma^2 \frac{\sigma^2_{01}}{\sigma^2_{00}} - 1 \right) \\
&\qquad + (q_{00} + q_{10}) \log \frac{1}{\gamma} + q_{00} \log \frac{\sigma_{00}}{\sigma_{01}} + q_{10} \log \frac{\sigma_{10}}{\sigma_{11}}.
\end{aligned}$$

This is a convex objective in γ (because the second derivative is non-negative) and by equating the derivative to zero, we have that the optimal γ^* must satisfy

$$\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)^{-1} (\mu_{01} - \mu_{11}) \left(\gamma^* (\mu_{01} - \mu_{11}) - (\mu_{00} - \mu_{10}) \right) + \gamma^* \left(q_{00} \frac{\sigma^2_{01}}{\sigma^2_{00}} + q_{10} \frac{\sigma^2_{11}}{\sigma^2_{10}} \right) - \frac{q_{00} + q_{10}}{\gamma^*} = 0.$$

Multiplying with $\gamma^* \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)$, we can write it as a quadratic equation as follows.

$$\begin{aligned}
&\left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right) \gamma^{*^2} - (\mu_{01} - \mu_{11})(\mu_{00} - \mu_{10}) \gamma^* \\
&\qquad - (q_{00} + q_{10}) \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right) = 0
\end{aligned}$$

The discriminant of the above quadratic polynomial is non-negative because the leading coefficient is positive and the constant term is negative. So this polynomial has two real roots. Moreover, since the constant term is negative, it cannot have both positive or both negative roots. Its only non-negative root is the optimal solution $\gamma^* \in \mathbb{R}_{\geq 0}$ we want.

$$\begin{aligned}
\gamma^* = \frac{(\mu_{01} - \mu_{11})(\mu_{00} - \mu_{10}) + \sqrt{\Delta}}{2 \left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right)}, \text{ where } \Delta = (\mu_{01} - \mu_{11})^2 (\mu_{00} - \mu_{10})^2 \\
+ 4 \left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right) (q_{00} + q_{10}) \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right).
\end{aligned}$$

□

Another intervention we can follow is to change all the subgroups of the given distribution. However, a quick check through the proof of Theorem 4.1 shows that this will lead to a non-convex program. However, just like Corollary C.2, we can show a reasonable intervention for the univariate case, where we change all four subgroups and search over a non-convex function using line search over a fairly large grid size.

Proposition C.3. *(All subgroup change for Exact Fairness) Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$, such that $q_{10}/q_{00} = q_{11}/q_{01}$, and let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ be univariate normal distributions, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y = i, A = a$ to $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\sigma}^2_{ia})$. Then minimizing $D_{\mathrm{KL}}\left(\tilde{D}||D\right)$ as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the constraints in Proposition 3.3 leads to a non-convex program.*

Furthermore, let $\gamma^* = \underset{\gamma \in (0, \infty)}{\arg\min}\, \mathcal{L}^*_\gamma$ for some non-convex function of γ that is only dependent on the original distribution parameters. Then, all the new distribution parameters $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ can be expressed as a function of γ^* and the original distribution parameters μ_{ia} and σ_{ia}.

Proof. We consider the following optimization program

$$D_{\mathrm{KL}}\left(\tilde{D}||D\right) = \sum_{(i,a)} q_{ia} D_{\mathrm{KL}}\left(\tilde{P}_{ia}||P_{ia}\right)$$
$$= \sum_{(i,a)} q_{ia}\left(\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2} + \frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2} + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right)$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the constraints

$$\frac{\tilde{\mu}_{01}-\tilde{\mu}_{11}}{\tilde{\sigma}_{11}} = \frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\tilde{\sigma}_{10}} \quad \text{and} \quad \frac{\tilde{\sigma}_{11}}{\tilde{\sigma}_{01}} = \frac{\tilde{\sigma}_{10}}{\tilde{\sigma}_{00}} \quad \text{and} \quad \tilde{\sigma}_{ia} \geq 0, \text{ for all } (i,a).$$

Let's fix $\gamma \in \mathbb{R}_{\geq 0}$ and minimize

$$\mathcal{L}_\gamma = \sum_{(i,a)} q_{ia}\left(\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2} + \frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2} + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right)$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the following constraints

$$\frac{\tilde{\mu}_{01}-\tilde{\mu}_{11}}{\tilde{\mu}_{00}-\tilde{\mu}_{10}} = \frac{\tilde{\sigma}_{11}}{\tilde{\sigma}_{10}} = \frac{\tilde{\sigma}_{01}}{\tilde{\sigma}_{00}} = \gamma \quad \text{and} \quad \tilde{\sigma}_{ia} \geq 0, \text{ for all } (i,a).$$

Now the objective $\mathcal{L}_\gamma$ is convex and for a fixed $\gamma \in \mathbb{R}_{\geq 0}$, the constraints on are linear in $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$. Let's denote the optimal solution for a fixed $\gamma \in \mathbb{R}_{\geq 0}$ by $\mu^*_{ia}(\gamma)$ and $\sigma^*_{ia}(\gamma)$, for $i, a \in \{0,1\}$. To find this, we can split the above objective into parts that can be optimized separately as follows.

$$\text{minimize} \quad \sum_{(i,a)} q_{ia}\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2} \quad \text{subject to} \quad \tilde{\mu}_{01}-\tilde{\mu}_{11} = \gamma(\tilde{\mu}_{00}-\tilde{\mu}_{10}), \quad \text{and}$$

$$\text{minimize} \quad \sum_{(i,a)} q_{ia}\left(\frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2} + \log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right) \quad \text{subject to} \quad \tilde{\sigma}_{i1} = \gamma\tilde{\sigma}_{i0}, \text{and } \tilde{\sigma}_{ia} \geq 0, \text{ for all } (i,a).$$

For each $i \in \{0,1\}$, by substituting $\tilde{\sigma}_{i1} = \gamma\tilde{\sigma}_{i0}$, we need to optimize a function in only one variable $\tilde{\sigma}_{i0}$. The optimal solutions $\sigma^*_{ia}(\gamma)$ turn out to be

$$\sigma^*_{i0}(\gamma) = \sqrt{\frac{q_{i0}+q_{i1}}{\frac{q_{i0}}{\sigma_{i0}^2}+\frac{q_{i1}\gamma^2}{\sigma_{i1}^2}}} \quad \text{and} \quad \sigma^*_{i1}(\gamma) = \gamma\sqrt{\frac{q_{i0}+q_{i1}}{\frac{q_{i0}}{\sigma_{i0}^2}+\frac{q_{i1}\gamma^2}{\sigma_{i1}^2}}}, \quad \text{for } i \in \{0,1\},$$

Now let's find the optimal solutions $\mu^*_{ia}(\gamma)$. The gradient of the objective must be parallel to the linear constraint, so

$$\frac{q_{00}(\mu^*_{00}(\gamma)-\mu_{00})}{\sigma_{00}^2} = -\gamma\lambda, \quad \frac{q_{01}(\mu^*_{01}(\gamma)-\mu_{01})}{\sigma_{01}^2} = \lambda,$$
$$\frac{q_{10}(\mu^*_{10}(\gamma)-\mu_{10})}{\sigma_{10}^2} = \gamma\lambda, \quad \frac{q_{11}(\mu^*_{11}(\gamma)-\mu_{11})}{\sigma_{11}^2} = -\lambda,$$

for some $\lambda \in \mathbb{R}$, which gives

$$\mu^*_{00}(\gamma) = -\gamma\lambda\frac{\sigma_{00}^2}{q_{00}}+\mu_{00}, \quad \mu^*_{01}(\gamma) = \lambda\frac{\sigma_{01}^2}{q_{01}}+\mu_{01}, \quad \mu^*_{10}(\gamma) = \gamma\lambda\frac{\sigma_{10}^2}{q_{10}}+\mu_{10}, \quad \mu^*_{11}(\gamma) = -\lambda\frac{\sigma_{11}^2}{q_{11}}+\mu_{11}.$$

Since $\mu^*_{ia}(\gamma)$ satisfies the constraint $\dfrac{\tilde{\mu}_{01}-\tilde{\mu}_{11}}{\tilde{\mu}_{00}-\tilde{\mu}_{10}} = \gamma$, we have

$$\frac{\lambda\frac{\sigma_{01}^2}{q_{01}}+\mu_{01}+\lambda\frac{\sigma_{11}^2}{q_{11}}-\mu_{11}}{-\gamma\lambda\frac{\sigma_{00}^2}{q_{00}}+\mu_{00}-\gamma\lambda\frac{\sigma_{10}^2}{q_{10}}-\mu_{10}} = \gamma, \quad \text{and hence,} \quad \lambda = \frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma_{01}^2}{q_{01}}+\frac{\sigma_{11}^2}{q_{11}}+\gamma^2\left(\frac{\sigma_{00}^2}{q_{00}}+\frac{\sigma_{10}^2}{q_{10}}\right)}.$$

Thus, we can express $\mu^*_{ia}(\gamma)$ as

$$\mu^*_{00}(\gamma) = -\gamma \frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{00}}{q_{00}}+\mu_{00}$$

$$\mu^*_{01}(\gamma) = \frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{01}}{q_{01}}+\mu_{01}$$

$$\mu^*_{10}(\gamma) = \gamma \frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{10}}{q_{10}}+\mu_{10}$$

$$\mu^*_{11}(\gamma) = -\frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{11}}{q_{11}}+\mu_{11}.$$

Thus, the optimal value of $\mathcal{L}_\gamma$ for a fixed $\gamma \in \mathbb{R}_{\geq 0}$ is given by

$$\mathcal{L}^*_\gamma = \sum_{(i,a)} q_{ia}\left(\frac{(\mu^*_{ia}(\gamma)-\mu_{ia})^2}{2\sigma^2_{ia}}+\frac{\sigma^*_{ia}(\gamma)^2-\sigma^2_{ia}}{2\sigma^2_{ia}}+\log\frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)}\right).$$

Dividing the above expression into three parts, the first part evaluates to

$$\begin{aligned}
\sum_{(i,a)} q_{ia}\frac{(\mu^*_{ia}(\gamma)-\mu_{ia})^2}{2\sigma^2_{ia}} &= \frac{q_{00}}{2\sigma^2_{00}}\frac{\gamma^2\lambda^2\sigma^4_{00}}{q^2_{00}}+\frac{q_{01}}{2\sigma^2_{01}}\frac{\lambda^2\sigma^4_{01}}{q^2_{01}}+\frac{q_{10}}{2\sigma^2_{10}}\frac{\gamma^2\lambda^2\sigma^4_{10}}{q^2_{10}}+\frac{q_{11}}{2\sigma^2_{11}}\frac{\lambda^2\sigma^4_{11}}{q^2_{11}} \\
&= \frac{\gamma^2\lambda^2\sigma^2_{00}}{2q_{00}}+\frac{\lambda^2\sigma^2_{01}}{2q_{01}}+\frac{\gamma^2\lambda^2\sigma^2_{10}}{2q_{10}}+\frac{\lambda^2\sigma^2_{11}}{2q_{11}} \\
&= \frac{\lambda^2}{2}\left(\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)\right) \\
&= \frac{1}{2}\left(\frac{\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11})}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}\right)^2\left(\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)\right) \\
&= \frac{1}{2}\frac{(\gamma(\mu_{00}-\mu_{10})-(\mu_{01}-\mu_{11}))^2}{\frac{\sigma^2_{01}}{q_{01}}+\frac{\sigma^2_{11}}{q_{11}}+\gamma^2\left(\frac{\sigma^2_{00}}{q_{00}}+\frac{\sigma^2_{10}}{q_{10}}\right)}.
\end{aligned}$$

The second part evaluates to

$$\begin{aligned}
\sum_{(i,a)} q_{ia}\frac{\sigma^*_{ia}(\gamma)^2-\sigma^2_{ia}}{2\sigma^2_{ia}} &= \sum_{(i,a)}\frac{q_{ia}}{2}\left(\frac{\sigma^*_{ia}(\gamma)^2}{\sigma^2_{ia}}-1\right) \\
&= \sum_{i=0}^{1}\frac{q_{i0}}{2}\left(\frac{q_{i0}+q_{i1}}{\sigma^2_{i0}\left(\frac{q_{i0}}{\sigma^2_{i0}}+\frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}-1\right)+\sum_{i=0}^{1}\frac{q_{i1}}{2}\left(\frac{\gamma^2(q_{i0}+q_{i1})}{\sigma^2_{i1}\left(\frac{q_{i0}}{\sigma^2_{i0}}+\frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}-1\right) \\
&= \sum_{i=0}^{1}\frac{q_{i0}}{2}\frac{q_{i1}\left(1-\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0}+q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}}+\sum_{i=0}^{1}\frac{q_{i1}}{2}\frac{q_{i0}\left(\gamma^2-\frac{\sigma^2_{i1}}{\sigma^2_{i0}}\right)}{q_{i0}\frac{\sigma^2_{i1}}{\sigma^2_{i0}}+q_{i1}\gamma^2} \\
&= \sum_{i=0}^{1}\frac{q_{i0}}{2}\frac{q_{i1}\left(1-\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0}+q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}}+\sum_{i=0}^{1}\frac{q_{i1}}{2}\frac{q_{i0}\left(\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}-1\right)}{q_{i0}+q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}} \\
&= 0,
\end{aligned}$$

and the third part evaluates to

$$\begin{aligned}
\sum_{(i,a)} q_{ia} \log \frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)} &= \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\sigma^2_{i0}}{\sigma^*_{i0}(\gamma)^2} + \frac{q_{i1}}{2} \log \frac{\sigma^2_{i1}}{\sigma^*_{i1}(\gamma)^2} \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\sigma^2_{i0}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0}+q_{i1}} + \frac{q_{i1}}{2} \log \frac{\sigma^2_{i1}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}{\gamma^2(q_{i0}+q_{i1})} \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}}{\frac{q_{i0}}{q_{i1}} + 1} + \frac{q_{i1}}{2} \log \frac{\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}}{\gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\left(\frac{q_{i0}}{q_{i1}} + 1\right)} \\
&= \sum_{i=0}^{1} \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\right) - \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + 1\right) - q_{i1} \log \gamma - q_{i1} \log \frac{\sigma_{i0}}{\sigma_{i1}}.
\end{aligned}$$

Putting it all together

$$\begin{aligned}
\mathcal{L}^*_\gamma &= \sum_{(i,a)} q_{ia} \left(\frac{(\mu^*_{ia}(\gamma) - \mu_{ia})^2}{2\sigma^2_{ia}} + \frac{\sigma^*_{ia}(\gamma)^2 - \sigma^2_{ia}}{2\sigma^2_{ia}} + \log \frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)} \right) \\
&= \frac{1}{2} \frac{\left(\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})\right)^2}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)} + \sum_{i=0}^{1} \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\right) - \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + 1\right) \\
&\qquad - q_{i1} \log \gamma - q_{i1} \log \frac{\sigma_{i0}}{\sigma_{i1}}.
\end{aligned}$$

Minimizing $\mathcal{L}^*_\gamma$ leads to a non convex program. Since γ is the ratio between variances of the new subgroup distribution, for a practical solution, we can do a line search over $\gamma \in (0, B)$ for some $B < \infty$. □

A popular intervention in the fairness literature is to equalize the first moment of the two sensitive groups or the mean outcomes of two groups, also known as the Calders-Verwer gap Calders & Verwer (2010); Kamishima et al. (2012); Chen et al. (2019). We, therefore, also study an intervention where we only change the mean of the under-privileged group and try to match it with the mean of the privileged group. We can show that the resulting optimization program is convex.

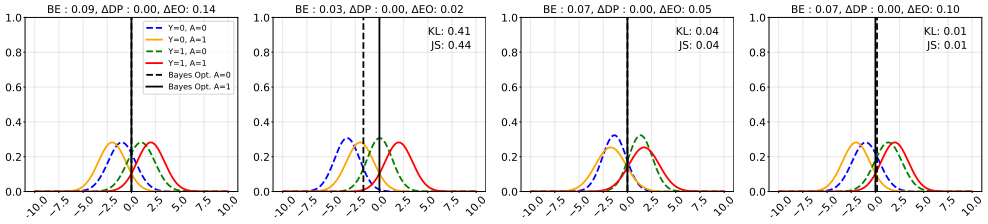
Proposition C.4. *(Affirmative Action by Equalizing First Moments) Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ be a univariate Normal distribution, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Then in the case of Affirmative mean change, where we impose the following constraints:*

$$\frac{q_{10}\,\tilde{\mu}_{10}}{q_{10}+q_{00}} + \frac{q_{00}\,\tilde{\mu}_{00}}{q_{10}+q_{00}} = \frac{q_{11}\,\mu_{11}}{q_{11}+q_{01}} + \frac{q_{01}\,\mu_{01}}{q_{11}+q_{01}},$$

we can efficiently minimize $D_{\mathrm{KL}}\left(\tilde{D}||D\right)$ as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$.

Proof. We are dealing with the following optimization problem:

$$\begin{aligned}
D_{\mathrm{KL}}\left(\tilde{D}||D\right) &= \sum_{i=0}^{1} q_{i0} D_{\mathrm{KL}}\left(\tilde{P}_{i0}||P_{i0}\right) \\
&= \sum_{i=0}^{1} q_{i0} \left(\frac{(\tilde{\mu}_{i0} - \mu_{i0})^2}{2\sigma^2_{i0}} + \frac{\tilde{\sigma}^2_{i0} - \sigma^2_{i0}}{2\sigma^2_{i0}} + \log \frac{\sigma_{i0}}{\tilde{\sigma}_{i0}} \right)
\end{aligned}$$



(a) Original Distribution (b) EF-Affirmative (c) EF-All Subgroups (d) Mean Matching

Figure 4: Comparison of Different Interventions when the subgroup distributions are shifted version of each other. While all methods achieve the same Bayes Error, Affirmative action is able to bring down the Bayes Error and achieve exact fairness.

as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ subject to the constraints

$$\frac{q_{10}}{q_{10}+q_{00}}\tilde{\mu}_{10} + \frac{q_{00}}{q_{10}+q_{00}}\tilde{\mu}_{00} = \frac{q_{11}}{q_{11}+q_{01}}\mu_{11} + \frac{q_{01}}{q_{11}+q_{01}}\mu_{01}$$

Since we are only changing the means and keeping the variances the same, the objective only depends on $\tilde{\mu}_{i0}$. Furthermore, let $K = {}^{(q_{10}+q_{00})}/_{(q_{11}+q_{01})} \cdot (q_{11}\mu_{11} + q_{01}\mu_{01})$ so that

$$\mathcal{L} = \sum_{i=0}^{1} q_{i0}\frac{(\tilde{\mu}_{i0}-\mu_{i0})^2}{2\sigma_{i0}^2}, \qquad \text{subject to } \tilde{\mu}_{00} = \frac{K-\tilde{\mu}_{10}}{q_{00}}.$$

Substituting the constraint on $\tilde{\mu}_{00}$ in the objective $\mathcal{L}$ gives us a convex quadratic in $\tilde{\mu}_{10}$, and the solution is obtained by setting the derivative to zero:

$$\tilde{\mu}_{00} = \frac{\frac{K}{\sigma_{10}^2\cdot q_{10}} - \frac{\mu_{10}}{\sigma_{10}^2} + \frac{\mu_{00}}{\sigma_{00}^2}}{\frac{q_{00}}{\sigma_{10}^2\cdot q_{10}} + \frac{1}{\sigma_{00}^2}}, \qquad \tilde{\mu}_{10} = \frac{\frac{K}{\sigma_{00}^2\cdot q_{00}} - \frac{\mu_{00}}{\sigma_{00}^2} + \frac{\mu_{10}}{\sigma_{10}^2}}{\frac{q_{10}}{\sigma_{00}^2\cdot q_{00}} + \frac{1}{\sigma_{10}^2}}, \qquad \tilde{\mu}_{01} = \mu_{01}, \qquad \text{and} \qquad \tilde{\mu}_{11} = \mu_{11}.$$

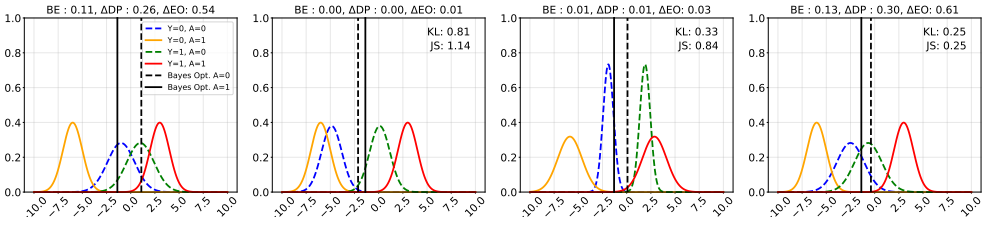
□

D Case Study Setup and Additional Univaraite Plots

We fix $q_{ia} \in (0,1)$ such that $q_{00}+q_{10}+q_{01}+q_{11}=1$, and our data generation works as follows. We simulate a data distribution where $Y=i, A=a$ with probability q_{ia} and $X \mid Y=i, A=a$ is sampled from a univariate Gaussian $\mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$. We choose homoskedastic Gaussians within each group $A=a$, i.e., $\sigma_{0a}=\sigma_{1a}$, so the we can show the Bayes optimal classifier boundary as a threshold. We choose different σ_{ia}'s that cover ground truth distribution that can the entire spectrum of being *ideal* or close to *ideal* to very far, and then we apply different interventions to change all or some subset of μ_{ia}'s and σ_{ia}'s to find the nearest *ideal* distribution in KL-divergence as given in Section 4.

We first look at a case where the subgroup distributions are the same shifted versions of each other in Figure 4. Note that all interventions, in this case, result in the same Bayes error (BE), but affirmative action brings the BE down with zero unfairness at the cost of incurring a deviation in terms of KL and JS divergence. However, in the next subplot, changing all four subgroups not only helps reduce the Bayes error and unfairness but also stays very close to the true distribution in the KL/JS sense. Matching the means also helps reduce the unfairness while staying close to the true distribution, but is sub-optimal compared to the EF-Affirmative and EF-All interventions.

Finally, in light of Proposition 3.2, we simulate the cost-sensitive risk for a different cost matrix C other than 0-1 loss by considering a threshold $t_C = {}^3/_4$ on $\eta(x,a)$ in Figure 5. The original distribution has high unfairness. EF-Affirmative intervention manages to achieve almost perfect fairness and zero error rate, but incurs relatively high KL/JS numbers. However, once again, changing all four subgroups, results in a solution that is perfectly fair and accurate, with low KL/JS. Mean Matching is unable to address the fairness-accuracy tension at all in this case and also manages to drift away from the true distribution, as indicated by non-zero KL/JS values.



(a) Original Distribution (b) EF-Affirmative (c) EF-All Subgroups (d) Mean Matching

Figure 5: Comparison of Different Interventions when we use a different threshold ($^3/_4$) than the Bayes optimal threshold ($^1/_2$). As derived in Proposition 3.2, the EF-Affirmative and EF-All interventions work with any threshold.