

---

# Supplementary Material: On Optimal Steering to Achieve Exact Fairness

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Proofs for Section 3: Ideal Distributions for Fair classification

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

3 **Lemma 1.1.** *Let  $\eta(x, a) = \Pr(Y = 1 | X = x, A = a)$ ,  $q_{ia} = \Pr(Y = i, A = a)$  and  $p_{ia}(x) =$   
 4  $\Pr(X = x | Y = i, A = a)$ . Then the Bayes optimal classifier can be written as  $h^*(x, a) =$   
 5  $\mathbb{I}\left(\log \frac{p_{1a}(x)}{p_{0a}(x)} \geq \log \frac{q_{0a}}{q_{1a}}\right)$ .*

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

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

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

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

11

□

12 We now prove Proposition 3.2 from the main paper.

13 **Proposition 1.2.** *(Proposition 3.2 in the main text) Let  $(X, Y, A)$  denote the features, class label,  
 14 and group membership, respectively, of a random data point from any data distribution  $D$  with  
 15  $q_{ia} = \Pr(Y = i, A = a)$ , for  $i \in \mathcal{Y}$  and  $a \in \mathcal{A}$ . Let  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$  be*

16 multivariate Normal distributions with mean  $\mu_{ia} \in \mathbb{R}^d$  and covariance matrix  $\Sigma_{ia} \in \mathbb{R}^{d \times d}$ , for  
 17  $i \in \mathcal{Y}$  and  $a \in \mathcal{A}$ . If the means  $\mu_{ia}$  and the covariance matrices  $\Sigma_{ia}$  satisfy

$$\begin{aligned} \Sigma_{ia}^{-1/2}(\mu_{ia} - \mu_{ja}) &= \Sigma_{ia'}^{-1/2}(\mu_{ia'} - \mu_{ja'}) \quad \text{and} \\ \Sigma_{ia}^{1/2} \Sigma_{ja}^{-1} \Sigma_{ia}^{1/2} &= \Sigma_{ia'}^{1/2} \Sigma_{ja'}^{-1} \Sigma_{ia'}^{1/2} \quad \text{and} \quad \frac{q_{ia}}{q_{ja}} = \frac{q_{ia'}}{q_{ja'}}, \quad \forall i, j \in \mathcal{Y}, a, a' \in \mathcal{A}, \end{aligned}$$

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

19 *Proof.* The Bayes optimal classifier for group  $A = a$  in a multi-class setting can be written down as  
 20 a maximum over posterior probabilities:

$$h^*(x, a) = \arg \max_{y \in \mathcal{Y}} \eta_y(x, a), \quad \text{where } \eta_y(x, a) = \Pr(Y = y | X = x, A = a).$$

21 We can say that  $h^*(x, a) = y$ , whenever the following happens:

$$h^*(x, a) = \arg \max_{y \in \mathcal{Y}} \eta_y(x, a) = \mathbb{I} \left( \frac{\eta_y(x, a)}{\eta_i(x, a)} \geq 1, \forall i \in \mathcal{Y} \right) = \mathbb{I} \left( \log \frac{p_{ya}(X)}{p_{ia}(X)} \geq \log \frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y} \right)$$

22 Using the above simplification, the EO-fairness condition  $\Pr(h^*(X, A) = y \mid Y = y, A = a) =$   
 23  $\Pr(h^*(X, A) = y \mid Y = 1, A = a') \quad \forall y \in \mathcal{Y}, a, a' \in \mathcal{A}$  means

$$\Pr \left( \log \frac{p_{ya}(X)}{p_{ia}(X)} \geq \log \frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y} \mid Y = y, A = a \right) = \Pr \left( \log \frac{p_{ya'}(X)}{p_{ia'}(X)} \geq \log \frac{q_{ia'}}{q_{ya'}}, \forall i \in \mathcal{Y} \mid Y = y, A = a' \right).$$

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

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

26 Now we can write

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

27 Let us denote the above expression as  $E_{yi}(r)$ . We can now write the group TPR as:

$$\Pr \left( \log \frac{p_{ya}(X)}{p_{ia}(X)} \geq \log \frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y} \mid Y = y, A = a \right) = \Pr \left( E_{yi}(R) \geq \log \frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y} \right),$$

28 for  $R \sim \mathcal{N}(\bar{0}, I_{d \times d})$ . Now if we have  $\frac{q_{ya}}{q_{ia}} = \frac{q_{ya'}}{q_{ia'}}$  and

$$\Sigma_{ya}^{-1/2}(\mu_{ya} - \mu_{ia}) = \Sigma_{ya'}^{-1/2}(\mu_{ya'} - \mu_{ia'}) \quad \text{and} \quad \Sigma_{ya}^{1/2} \Sigma_{ia}^{-1} \Sigma_{ya}^{1/2} = \Sigma_{ya'}^{1/2} \Sigma_{ia'}^{-1} \Sigma_{ya'}^{1/2},$$

29 then the probability of the above event written in terms  $R \sim \mathcal{N}(\bar{0}, I_{d \times d})$  becomes identical  $\forall a, a' \in$   
 30  $\mathcal{A}, i \in \mathcal{Y}$ . Hence, the Bayes optimal classifier satisfies equal opportunity with these set of conditions.

31 □

**Proposition 1.3.** (Proposition 3.3 in the main text) 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  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$  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

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

*Proof.* 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 [7] and [14]. The distribution  $D$  is *ideal* for equal opportunity if  $\Pr(\eta(X, A) \geq t \mid Y = i, A = 0) = \Pr(\eta(X, A) \geq t \mid Y = i, A = 1)$ , for all thresholds  $t \in [0, 1]$  and  $i \in \{0, 1\}$ . Since the CDFs are identical, the random variables  $\eta(X, A) \mid Y = i, A = 0$  and  $\eta(X, A) \mid Y = i, A = 1$  must be identical. Note that

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

If  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$ , then  $X \mid Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$ . Thus, for  $\eta(X, A) \mid Y = i, A = 0$  and  $\eta(X, A) \mid Y = i, A = 1$  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

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

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

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

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

## 49 2 Proofs for Section 4

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

52 **Lemma 2.1.** *Let  $(X, Y, A)$  denote the features, binary class label, and binary group membership,*  
 53 *respectively, of a random data point from any data distribution  $D$  with  $q_{ia} = \Pr(Y = i, A = a)$ , for*  
 54  *$i \in \mathcal{Y}$  and  $a \in \mathcal{A}$ . Let  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$  be multivariate Normal distributions with*  
 55 *mean  $\mu_{ia} \in \mathbb{R}^d$  and covariance matrix  $\Sigma_{ia} \in \mathbb{R}^{d \times d}$ . Let  $\tilde{D}$  denote a distribution obtained by keeping*  
 56  *$(Y, A)$  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,*

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

*Proof.*

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

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

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

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

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

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

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

61

□

62 **Theorem 2.2.** (Theorem 4.1 in the main text) Let  $(X, Y, A)$  denote the features, binary class label,  
63 and binary group membership, respectively, of a random data point from any data distribution  $D$  with  
64  $q_{ia} = \Pr(Y = i, A = a)$ , for  $i \in \{0, 1\}$  and  $a \in \{0, 1\}$ , such that  $q_{10}/q_{00} = q_{11}/q_{01}$ . Let  $X|Y =$   
65  $i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$  be multivariate Normal distributions, with mean  $\mu_{ia} \in \mathbb{R}^d$  and covariance  
66 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  
67  $(Y, A)$  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  
68 the case of Affirmative action (changing only  $\tilde{\mu}_{i0}$  and  $\tilde{\Sigma}_{i0}$ ), we can efficiently minimize  $D_{\text{KL}}(\tilde{D} || D)$   
69 as a function of the variables  $\tilde{\mu}_{i0}$  and  $\tilde{\Sigma}_{i0}$  subject to the constraints in Proposition 1.2, so that the  
70 Bayes optimal classifier on the optimal  $\tilde{D}$  is guaranteed to be EO-fair.

71 *Proof.* Using Lemma 2.1 and Proposition 1.2, our objective is to minimize

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

72 subject to the constraints

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

73 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  $\tilde{\Sigma}_{i0}^{1/2} = \Gamma \tilde{\Sigma}_{i1}^{1/2}$ )

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

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

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  $x^T A x + x^T b$  subject to  $x^T c = e$  [8].

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  $\tilde{\mu}_{00}$ ),

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

This is a convex objective in  $\tilde{\mu}_{00}$  because its Hessian is positive semidefinite, i.e.,  $q_{00} \Sigma_{00}^{-1} + q_{10} \Sigma_{10}^{-1} \succcurlyeq 0$  [3]. 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

$$\begin{aligned} \mu_{00}^*(\Gamma) &= (q_{00} \Sigma_{00}^{-1} + q_{10} \Sigma_{10}^{-1})^{-1} (\Sigma_{00}^{-1} \mu_{00} + \Sigma_{10}^{-1} \mu_{10} - \Sigma_{10}^{-1} \Gamma(\mu_{11} - \mu_{01})) \quad \text{and} \\ \mu_{10}^*(\Gamma) &= (q_{00} \Sigma_{00}^{-1} + q_{10} \Sigma_{10}^{-1})^{-1} (\Sigma_{00}^{-1} \mu_{00} + \Sigma_{10}^{-1} \mu_{10} - \Sigma_{10}^{-1} \Gamma(\mu_{11} - \mu_{01})) + \Gamma(\mu_{11} - \mu_{01}) \\ \Sigma_{00}^*(\Gamma) &= (\Gamma \Sigma_{01}^{1/2})^2 \\ \Sigma_{10}^*(\Gamma) &= (\Gamma \Sigma_{11}^{1/2})^2. \end{aligned}$$

By substituting these, when we look at the objective as a function of a positive semidefinite matrix-valued variable  $\Gamma$ , it turns out to be convex. This requires rewriting the expressions using the identities  $\text{tr}(AB) = \text{tr}(BA)$ ,  $\det(AB) = \det(A)\det(B)$ , and most importantly,  $\text{tr}(AXBX) = \text{tr}((A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T)$  and  $\log \det(AXBX) = \log \det(A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T$ , for symmetric, positive semidefinite matrices  $A, B, X$  [12]. The convexity of the objective in  $\Gamma$  follows from the convexity of  $\text{tr}(AXBX)$  and  $-\log \det(X)$  for matrix-valued variable  $X$ . Finally, we can solve it efficiently to get the optimal  $\Gamma^*$ .  $\square$

**Corollary 2.3.** (Corollary 4.2 in the main text) For the case where  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$  are univariate normal distributions, for  $i, a \in \{0, 1\}$ , the optimal distribution  $\tilde{D}$  from Theorem 4.1, with  $\gamma^*$  being a function of the original distribution parameters, can be written down as:

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

*Proof.*

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

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

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

101 respectively. Hence,

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

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

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

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

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

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

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

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

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

107 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  $\tilde{\sigma}_{i0}$ .  
 108 Let's denote the optimal solution for a fixed  $\gamma \in \mathbb{R}_{\geq 0}$  by  $\mu_{i0}^*(\gamma)$  and  $\sigma_{i0}^*(\gamma)$ , for  $i \in \{0, 1\}$ . For  
 109 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  
 110  $\tilde{\mu}_{00} = \tilde{\mu}_{10} + \gamma(\mu_{01} - \mu_{11})$ , we only need to minimize the following convex, quadratic objective in a  
 111 single variable  $\tilde{\mu}_{10}$ ,

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

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

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

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

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

114 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

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

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

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

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

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



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

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

$$\Delta = (\mu_{01} - \mu_{11})^2(\mu_{00} - \mu_{10})^2$$

$$+ 4 \left( (\mu_{01} - \mu_{11})^2 + \sigma_{01}^2 + \sigma_{11}^2 + \frac{q_{10}\sigma_{00}^2}{q_{00}\sigma_{10}^2}\sigma_{11}^2 + \frac{q_{00}\sigma_{10}^2}{q_{10}\sigma_{00}^2}\sigma_{01}^2 \right) (q_{00} + q_{10}) \left( \frac{\sigma_{00}^2}{q_{00}} + \frac{\sigma_{10}^2}{q_{10}} \right).$$

122

□

123 **Proposition 2.4.** (Proof of Proposition 4.3 in the main text) Let  $(X, Y, A)$  denote the features,  
 124 binary class label, and binary group membership, respectively, of a random data point from any  
 125 data distribution  $D$  with  $q_{ia} = \Pr(Y = i, A = a)$ , for  $i \in \{0, 1\}$  and  $a \in \{0, 1\}$ , such that  
 126  $q_{10}/q_{00} = q_{11}/q_{01}$ , and let  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$  be univariate normal distributions, for  
 127  $i \in \{0, 1\}$  and  $a \in \{0, 1\}$ . Let  $\tilde{D}$  denote a distribution obtained by keeping  $(Y, A)$  unchanged and  
 128 only changing  $X|Y = i, A = a$  to  $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\sigma}_{ia}^2)$ . Then minimizing  $D_{\text{KL}}(\tilde{D}||D)$   
 129 as a function of the variables  $\tilde{\mu}_{ia}$  and  $\tilde{\sigma}_{ia}$  subject to the constraints in Proposition 3.2 leads to a  
 130 non-convex program. Furthermore, let  $\gamma^* = \arg \min_{\gamma \in (0, \infty)} \mathcal{L}_\gamma^*$  for some non-convex function of  $\gamma$  that is  
 131 only dependent on the original distribution parameters. Then, all the new distribution parameters  $\tilde{\mu}_{ia}$   
 132 and  $\tilde{\sigma}_{ia}$  can be expressed as a function of  $\gamma^*$  and the original distribution parameters  $\mu_{ia}$  and  $\sigma_{ia}$ .

133 *Proof.* We consider the following optimization program

$$D_{\text{KL}}(\tilde{D}||D) = \sum_{(i,a)} q_{ia} D_{\text{KL}}(\tilde{P}_{ia}||P_{ia})$$

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

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

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

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

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

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

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

137 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  $\tilde{\sigma}_{ia}$ .  
 138 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  
 139 find this, we can split the above objective into parts that can be optimized separately as follows.

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

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

140 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  
 141  $\tilde{\sigma}_{i0}$ . The optimal solutions  $\sigma_{ia}^*(\gamma)$  turn out to be

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

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

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

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

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

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

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

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

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

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

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

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

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

149 The second part evaluates to

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

150 and the third part evaluates to

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

151 Putting it all together

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

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

155 **Bound on Unfairness and Error Rate** For completeness, we now derive upper bounds on the error  
156 rate and the unfairness gap  $\Delta_{EO}$  of the Bayes optimal classifier  $\tilde{h}$  on  $\tilde{D}$  with respect to the original  
157 distribution  $D$ . These bounds show that both the accuracy loss and the fairness gap depend only  
158 on the KL divergence between  $D$  and  $\tilde{D}$ . It also shows that the optimal value of our optimization  
159 problem can be used to approximately translate the accuracy guarantee of  $\tilde{h}$  from  $\tilde{D}$  to  $D$ .

160 **Proposition 2.5.** *Let  $\text{err}(h, D)$  denote the error rate (expected 0-1 loss) of a classifier  $h$  on the*  
161 *distribution  $D$ . Let  $d_{TV}(\tilde{D}, D)$  denote the total variation distance between two distributions  $\tilde{D}$  and*  
162  *$D$ , while  $D_{KL}$  denotes the KL-Divergence between them. Denote the Bayes optimal classifier on the*  
163 *ideal distribution  $\tilde{D}$  as  $\tilde{h}$  (and similarly the Bayes optimal classifier  $h$ ). Then, we can bound the error*  
164 *rate and Equal opportunity of  $\tilde{h}$  on the original distribution  $D$  as follows:*

$$|\text{err}(\tilde{h}, D) - \text{err}(\tilde{h}, \tilde{D})| \leq \sqrt{2D_{KL}(\tilde{D}, D)} \quad \text{and} \quad \Delta_{EO}(\tilde{h}, D) \leq \sqrt{8D_{KL}(\tilde{D}||D)}.$$

165 *Proof.* For the sake of this proof, we assume a countable data domain. Using the definition of the  
166 expected 0 – 1 loss, we can write:

$$\begin{aligned}
&\text{err}(\tilde{h}, D) - \text{err}(\tilde{h}, \tilde{D}) \\
&= \sum_{(x,y,a)} \mathbb{I}(\tilde{h}(x, a) \neq y) \cdot (p(x, y, a) - \tilde{p}(x, y, a)) \\
&\leq 2d_{TV}(\tilde{D}, D) \leq \sqrt{2D_{KL}(\tilde{D}, D)} \quad (\text{Pinsker's Inequality [5]}).
\end{aligned}$$

167 The first inequality follows from writing the error as expected 0-1 loss and using the definition of  
168 total variation distance. The last line follows from Pinsker's inequality [5]. We can similarly prove  
169 the other direction to obtain the first inequality. Similarly, for the true positive rate of group  $a$ ,  
170  $TPR_a(\tilde{h}, D) - TPR_a(\tilde{h}, \tilde{D}) \leq 2d_{TV}(\tilde{D}, D)$ .

171 We can also write for the other group  $A = a'$ ,  $TPR_{a'}(\tilde{h}, \tilde{D}) - TPR_{a'}(\tilde{h}, D) \leq 2d_{TV}(\tilde{D}, D)$ .  
172 Adding both LHS and RHS and repeating the exercise in the other direction, noting that the TPR  
173 difference of  $\tilde{h}$  in  $\tilde{D}$  is zero (since in the ideal distribution we have exact fairness), we can bound the  
174 absolute value of the TPR difference, which is our definition of  $\Delta_{EO}$ , we get:

$$\Delta_{EO}(\tilde{h}, D) \leq \sqrt{8D_{KL}(\tilde{D}||D)}$$

175  $\square$

176 **Equalizing the first moment** A popular intervention in the fairness literature is to equalize the  
177 first moment of the two sensitive groups or the mean outcomes of two groups, also known as the  
178 Calders-Verwer gap [4, 11, 6]. We, therefore, also study an intervention where we only change the  
179 mean of the under-privileged group and try to match it with the mean of the privileged group. We can  
180 show that the resulting optimization program is convex. We leverage this intervention in Section 5 of  
181 the main text.

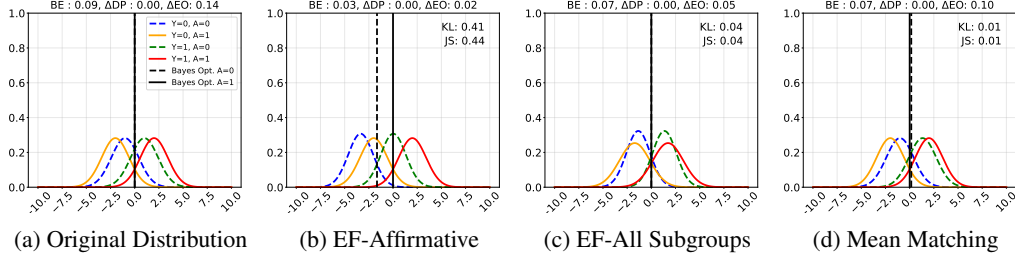


Figure 1: Comparison of Different Interventions when the subgroup distributions are shifted version of each other. While all methods achieve the same Bayes Error, Affirmative action is able to bring down the Bayes Error and achieve exact fairness.

**Proposition 2.6.** (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  $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$  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  $D_{\text{KL}}(\tilde{D}||D)$  as a function of the variables  $\tilde{\mu}_{i0}$  and  $\tilde{\Sigma}_{i0}$ .

*Proof.* We are dealing with the following optimization problem:

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

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

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

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

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

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:

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

□

### 3 Additional Figures for Section 5

In this section, we lay out additional plots from our Gaussian case study. We first describe the setup. We modify a stylized setting of Gaussian distributions from previous work (see Definition 3.1 in [13], Section 5.3 in [1]) to investigate the unfairness and the Bayes optimal error on the

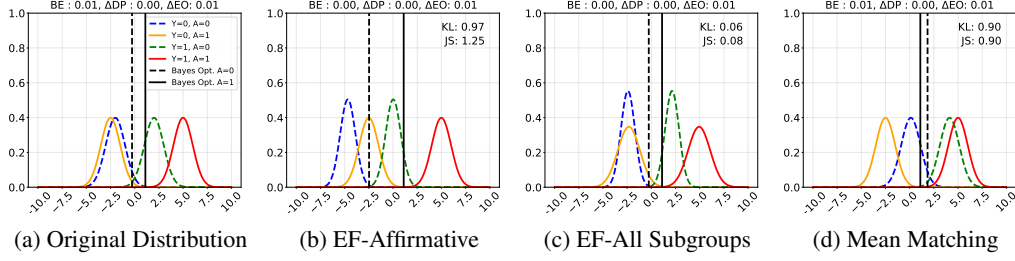


Figure 2: Comparison of Different Interventions when the original distribution is already fair. In this case, EF-All ensures that it stays close to the true distribution, as no intervention as required, while others relatively deviate.

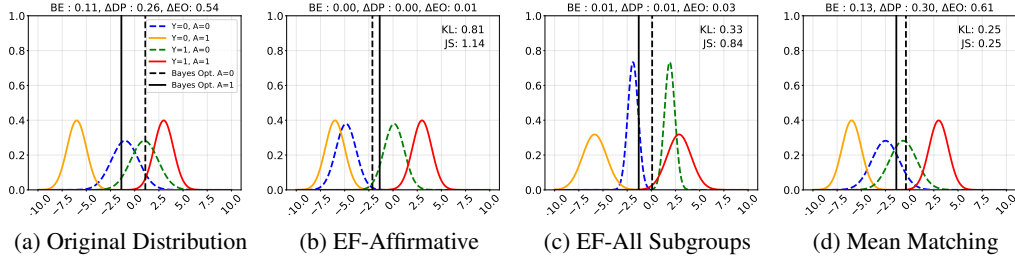


Figure 3: Comparison of Different Interventions when we use a different threshold ( $3/4$ ) than the Bayes optimal threshold ( $1/2$ ). As derived in Proposition 1.3, the EF-Affirmative and EF-All interventions work with any threshold.

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

207 We first look at a case where the subgroup distributions are the same shifted versions of each other in  
208 Figure 1. Note that all interventions, in this case, result in the same Bayes error (BE), but affirmative  
209 action brings the BE down with zero unfairness at the cost of incurring a deviation in terms of KL  
210 and JS divergence. However, in the next subplot, changing all four subgroups not only helps reduce  
211 the Bayes error and unfairness but also stays very close to the true distribution in the KL/JS sense.  
212 Matching the means also helps reduce the unfairness while staying close to the true distribution, but  
213 is sub-optimal compared to the EF-Affirmative and EF-All interventions.

214 Next, we look at a case where the Bayes optimal classifier is already fair ( $\Delta EO$  is close to 0 while  
215  $\Delta DP=0$ ) in Figure 2. The expected solution here should be that any intervention must leave the  
216 distribution as it is. EF-Affirmative intervention keeps the unfairness and error rate numbers as it is,  
217 but deviates from the true distribution, as indicated by the KL/JS divergences. However, the EF-All  
218 intervention only makes major changes to variances and stays close to the true distribution. The  
219 Mean Matching intervention shifts both the under-privileged subgroups and strays away from the  
220 true distribution, as indicated by relatively high KL/JS values.

221 Finally, in light of Proposition 1.3, we simulate the cost-sensitive risk for a different cost matrix  
222  $C$  other than 0-1 loss by considering a threshold  $t_C = 3/4$  on  $\eta(x, a)$  in Figure 3. The original  
223 distribution has high unfairness. EF-Affirmative intervention manages to achieve almost perfect  
224 fairness and zero error rate, but incurs relatively high KL/JS numbers. However, once again, changing  
225 all four subgroups, results in a solution that is perfectly fair and accurate, with low KL/JS. Mean  
226 Matching is unable to address the fairness-accuracy tension at all in this case and also manages to  
227 drift away from the true distribution, as indicated by non-zero KL/JS values.

## 4 Details and Additional Results for Section 6

In this section, we lay down all the details for the experiments performed for LLM steering. The code to reproduce our results is provided *here*. For the multi-class experiments, we use a lot of helper functions from the code of Singh et al. [15]<sup>1</sup>. For the emotion steering experiments, we reproduce the methodology from Zhao et al. [16] and provide the Jupyter notebook in our code.

### 4.1 Reducing Disparity in Multi-class classification

To apply our intervention for multi-class settings, we first come up with a version of Theorem 4.1 for multiple classes. We show this for a univariate distribution, and for our intervention, we assume diagonal covariance. Since our experiment setup only requires two groups for each class, we show the effective constraints assuming two sensitive groups, but this methodology can be readily extended to handle a countable number of groups as well. To make our program convex (and affirmative), we fix a class  $y \in \mathcal{Y}$ . We fix our class  $y^*$  according to the following heuristic:  $y^* = \arg \min_{y \in \mathcal{Y}} \Delta_y TPR(\hat{h})$ ,

where  $\hat{h}$  is the empirical risk minimizer on the given data. This fixes our ratio  $\gamma_\sigma = \frac{\sigma_{y^*1}}{\sigma_{y^*0}}$  and  $\gamma_q = \frac{q_{y^*1}}{q_{y^*0}}$ .

We can now write a multi-class version of the optimization program in Theorem 4.1:

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

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

$$\frac{\tilde{\mu}_{i1} - \tilde{\mu}_{j1}}{\tilde{\mu}_{i0} - \tilde{\mu}_{j0}} = \frac{\tilde{\sigma}_{i1}}{\tilde{\sigma}_{i0}} = \frac{\tilde{\sigma}_{j1}}{\tilde{\sigma}_{j0}} = \gamma_\sigma, \quad \frac{q_{i1}}{q_{i0}} = \frac{q_{j1}}{q_{j0}} = \gamma_q \quad \text{and} \quad \tilde{\sigma}_{ia} \geq 0, \text{ for all } i \in \mathcal{Y}, j \in \mathcal{Y} \setminus \{i\}.$$

Just like in the proof of Theorem 4.1, the resulting program will result in separable objectives for a class  $y$  and then in the underlying optimization variables  $\tilde{\mu}_{ya}$  and  $\tilde{\sigma}_{ya}$ :

**Program for  $\tilde{\mu}_{ya}$ :**

$$q_{y0} \frac{(\tilde{\mu}_{y0} - \mu_{y0})^2}{2\sigma_{y0}^2} + q_{y1} \frac{(\tilde{\mu}_{y1} - \mu_{y1})^2}{2\sigma_{y1}^2}, \text{ subject to } \tilde{\mu}_{y1} - \tilde{\mu}_{y^*1} = \gamma(\tilde{\mu}_{y0} - \tilde{\mu}_{y^*0})$$

**Program for  $\tilde{\sigma}_{ya}$ :**

$$q_{y0} \left( \frac{\tilde{\sigma}_{y0}^2 - \sigma_{y0}^2}{2\sigma_{y0}^2} + \log \frac{\sigma_{y0}}{\tilde{\sigma}_{y0}} \right) + q_{y1} \left( \frac{\tilde{\sigma}_{y1}^2 - \sigma_{y1}^2}{2\sigma_{y1}^2} + \log \frac{\sigma_{y1}}{\tilde{\sigma}_{y1}} \right), \text{ subject to } \tilde{\sigma}_{y1} = \gamma \tilde{\sigma}_{y0},$$

where  $y^*$  is the class we fixed earlier and  $\gamma = \gamma_\sigma$ . The solution for the following programs are the following:

$$\tilde{\sigma}_{y0} = \pm \sqrt{\frac{q_{y0} + q_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2} + \frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}}, \quad \tilde{\sigma}_{y1} = \pm \gamma \sqrt{\frac{q_{y0} + q_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2} + \frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}}, \quad \tilde{\mu}_{y0} = \frac{\frac{q_{y0}}{\sigma_{y0}^2} \mu_{y0} + \frac{\gamma q_{y1}}{\sigma_{y1}^2} (\mu_{y1} - \mu_{y^*1} + \gamma \mu_{y^*0})}{\frac{q_{y0}}{\sigma_{y0}^2} + \frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}, \text{ and } \tilde{\mu}_{y1} = \frac{\frac{q_{y0}}{\sigma_{y0}^2} (\mu_{y^*1} - \gamma \mu_{y^*0} + \gamma \mu_{y0}) + \frac{\gamma^2 q_{y1}}{\sigma_{y1}^2} \mu_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2} + \frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}.$$

Once we have the corrected distributions  $\mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$ , we set up an affine intervention, following the design choices of Singh et al. [15]. We assume an affine relationship between the original and transformed samples per subgroup:  $Y = a_{ya}X + b_{ya}$ , where  $Y \sim \mathcal{N}(\tilde{\mu}_{ya}, \tilde{\sigma}_{ya})$  and  $X \sim \mathcal{N}(\mu_{ya}, \sigma_{ya})$ . Taking expectation on both sides gives us:  $\tilde{\mu}_{ya} = a_{ya}\mu_{ya} + b_{ya}$ ,  $\tilde{\sigma}_{ya}^2 = a_{ya}^2\sigma_{ya}^2$ , and we get the following coefficients:  $a_{ya} = \pm \frac{\tilde{\sigma}_{ya}}{\sigma_{ya}}$  and  $b_{ya} = \tilde{\mu}_{ya} \pm \frac{\tilde{\sigma}_{ya}}{\sigma_{ya}}\mu_{ya}$ .

<sup>1</sup><https://github.com/shauli-ravfogel/affine-steering>

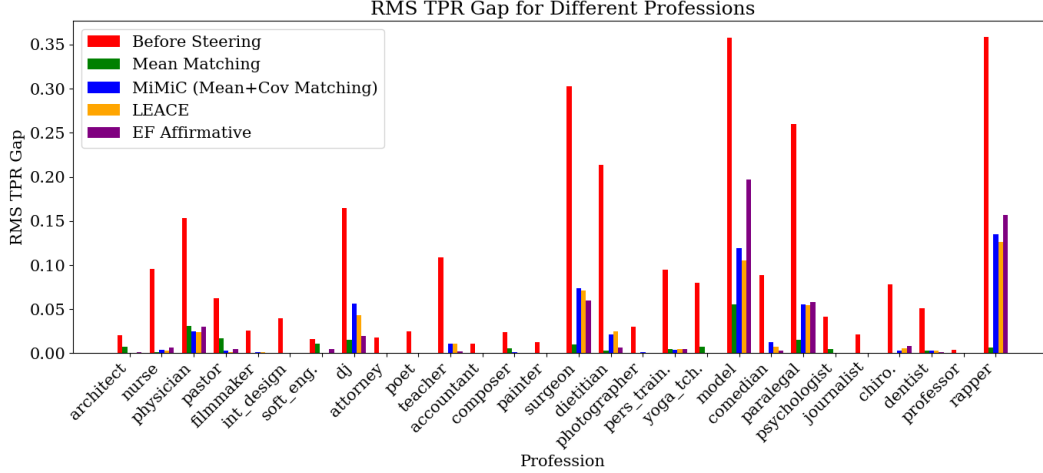


Figure 4: TPR-gap between Gender groups for all professions. All methods to steer feature representations achieve roughly the same accuracy (in the range of 0.77-0.79). Our intervention (EF Affirmative) is able to significantly reduce the TPR-gap for all professions. In many cases, it is even comparable or better than previous interventions Belrose et al. [2], Singh et al. [15].

257 We have two choices of the parameters corresponding to the positive and negative solutions. Since we  
 258 are working with empirical estimates, we use the validation error to decide the best set of estimates.  
 259 A detailed implementation is given in the code. To implement the conditions for  $q_{ya}$ , we use the  
 260 reweighing scheme of Kamiran and Calders [10]. The plot in the main text (Figure 3) assumes  
 261  $q_{y0} = q_{y1}$  for the corrected covariances. Figure 4 shows the plot with no such assumption for  $q_{ya}$   
 262 and confirms with same trends as observed in Figure 3.

## 263 4.2 Steering activations for Joyful generation

264 Zhao et al. [16] propose to obtain a distribution over the steering vectors for a concept instead of  
 265 a single steering vector. In this section, we lay out all our prompts and the design choices for the  
 266 emotion steering pipeline.

267 We first generate training data for each concept (joyful, angry) for each of the groups (horror, comedy),  
 268 resulting in four subgroups. The following prompt was used to generate an initial set of data:

Prompt to generate 1000 Comedy movie reviews that are joyful.

Compose a concise 30-word movie review, assuming it is a comedy movie, that covers these four aspects: plot, sound and music, cultural impact, and emotional resonance. Choose a joyful tone for your review. For the plot, comment on its structure or originality. Regarding sound and music, mention how it enhances the storytelling. For cultural impact, touch on any relevant social commentary. Finally, describe how the film resonates emotionally. Ensure your joyful tone is consistent throughout the review. Please include emotions like “joyful” in these texts and generate 1000 samples.

269

270 We obtain the last token embeddings from each layer of a Llama-3.1 8B model [9] for each of the  
 271 samples. We now proceed towards obtaining the steering vectors. We want to obtain a steering vector  
 272 for each group. Treating angry reviews as an irrelevant sample for the ‘joyful’ concept, we assign  
 273  $y = 0$  to angry samples and  $y = 1$  to joyful samples. Because we want to estimate a distribution over  
 274 the steering vectors instead of a single vector, we sample 300 points with replacement and repeat this  
 275 for 50 iterations. In each of these iterations, we train a Logistic Regression model to classify between  
 276 the relevant and the irrelevant samples. We get 50 weight vectors using this pipeline, and we use  
 277 those to obtain the sample mean and covariance. We denote the resulting distribution for the steering  
 278 vectors for layer  $l$  as  $\mathcal{N}(\mu_{1a}^l, \Sigma_{1a}^l)$  where  $1a$  denotes that this distribution represents the steering  
 279 vector for joyful emotion for a group  $A = a$  (horror or comedy reviews). To apply our intervention



280 later, we also obtain the steering vector in the other direction by flipping the relevance labels, i.e.  
 281 joyful  $\rightarrow$  angry, and we denote the resulting Gaussian distribution with  $\mathcal{N}(\mu_{0a}^l, \Sigma_{0a}^l)$ .  
 282 To perform steering, we now sample a steering vector  $v_c^l \sim \mathcal{N}(\mu_{1a}^l, \Sigma_{1a}^l)$  and add it to the last token  
 283 representation of layer  $l$  with strength  $a$ :  $h^l = (1 - a)h^l + av_c^l$ . To measure the performance of  
 284 steering, we ask the Llama model to generate angry reviews using the following prompt:

Prompt to evaluate the joyfulness of generated reviews (for group horror).

**System Prompt:** You are a well informed movie critic.

**User Prompt:** Compose a concise 10-word movie review, assuming it is a horror movie. The tone of the review must be angry. The review should be within 10 words. Generate 100 samples. Please make sure that you do not repeat the same review and keep each review under 10 words. Also, make sure that the reviews are coherent and not repetitive.

285

286 We first measure the joyfulness score of the generated reviews without adding any steering vectors.  
 287 Just like Zhao et al. [16], we use the following prompt to ask the GPT 4.1 model to rate the joyfulness  
 288 of the text:

Prompt to generate angry 10-word reviews for the horror genre.

**System Prompt:** You are a deterministic evaluation agent trained to assess text samples using two strictly defined criteria: (1) emotional tone (joyful score) and (2) linguistic structure (chaotic score).

For each input, your task is to produce exactly two integer scores between 0 and 10. These must follow the user's scoring rubric precisely. The first score reflects emotional positivity or dark exuberance without considering sentence structure. The second score reflects linguistic disorder or repetition, without considering sentiment.

You must produce output in CSV format, strictly matching the following structure: "sample", joyful score, chaotic score

Do not include any additional content, commentary, or variation. Your outputs must be fully deterministic: identical input always produces identical output.

Your responses should match the tone and evaluation logic demonstrated in the provided reference examples.

**User Prompt:** You are an automated evaluator tasked with rating horror movie reviews using two fixed criteria: **joyful expression** and **textual coherence and novelty**.

For each review, you must generate exactly two integer scores as described below:

### **1. Joyful Score**

Rate from 0 to 10 based strictly on emotional tone:

- \* **0** = not joyful at all (strongly negative, angry, or emotionally flat tone)
- \* **10** = extremely joyful, enthusiastic, or darkly exuberant in tone
- \* Intermediate values reflect a mix of enthusiasm and criticism
- \* Focus solely on affect and sentiment; ignore linguistic structure for this score

### **2. Chaotic Score**

Rate from 0 to 10 based strictly on linguistic organisation:

- \* **0** = fully coherent, fluent, and structurally sound
- \* **10** = highly chaotic, incoherent, repetitive, or nonsensical
- \* Intermediate values reflect partial disorganisation, verbosity, or broken phrasing
- \* Focus only on syntax, flow, and novelty of expression; ignore emotional content

### **Reference Examples**

**Joyful and Coherent Horror Reviews:**

- \* "Blood-soaked fun ensues in this delightfully terrifying slasher film."
- \* "Chilling thrills abound in this creepy haunted mansion tale."
- \* "Jump scares galore in this electrifying horror comedy gem."
- \* "Unsettling unease fills this unnerving psychological horror masterpiece."
- \* "Bone-chilling chills chill to the bone in this one."

**Angry and Coherent Horror Reviews:**

- \* "Abysmal plot twists ruined what could've been a decent film."
- \* "Mind-numbing terror fails to deliver in this lazy horror."
- \* "Weak jump scares can't save this trainwreck disaster."
- \* "Poor production values ruin what little suspense exists."
- \* "Frustratingly predictable, making it boring and unscary too."

### **Output Format**

\* For each sample, return one line in strict CSV format: "sample", joyful score, chaotic score

**Example Output:**

"sample, joyful score, chaotic score

"This horror film was painfully dull and predictable.", joyful\_score\_1, chaotic\_score\_1

"Terrifying, stylish, and packed with chilling moments!", joyful\_score\_2, chaotic\_score\_2

"

\* Do **not** include explanations, commentary, or additional formatting.

\* Output must be **fully deterministic**: the same input must always yield the same scores.

Begin processing the dataset now. Here is the batch of review samples:

A few notes on evaluation are in order. Zhao et al. [16] report both joyfulness and coherence scores for the generated text. However, we observed that coherence scores were all over the place and did not make sense. Second, Zhao et al. evaluate using the GPT-4o model, whereas we used the GPT 4.1 model since we observed that the joyful scores corroborated more with the qualitative inspection of the generated samples.

Following the above pipeline, we observe an increase in joyfulness scores of the generated reviews by a Llama model after steering. However, since the effectiveness of joyful steering was not the same for the horror and comedy movie review generations, we apply our affirmative intervention (Theorem 4.1), assuming that the horror and comedy movie reviews define two groups. Let the modified steering vector be denoted by  $\tilde{v}_c^l \sim \mathcal{N}(\tilde{\mu}_{1a}^l, \tilde{\Sigma}_{1a}^l)$ , where the new gaussian distribution is obtained after applying the affirmative action intervention from Theorem 4.1 assuming horror group is the under-privileged group.

However, simply replacing  $v_c^l$  will not work. We demonstrate that empirically in the main text, where in Figure 4,  $\alpha = 1$  corresponds to using  $\tilde{v}_c^l$  instead of  $v_c^l$ . But we can always use  $\tilde{v}_c^l$  to nudge the existing steering vector  $v_c^l$  in the right direction. To do that, we modify the steering vector and the representation  $h^l$  with the following rule:  $h^l = (1 - \alpha)h^l + \alpha((1 - \alpha)v_c^l + \alpha\tilde{v}_c^l)$ , where  $\alpha$  controls the strength of mixing the old and new steering vectors. In Figure 4, we show that for small values of  $\alpha$ , the steering vector indeed starts performing better in steering the reviews of the horror group towards a more joyful tone.

## References

- [1] 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 Vredenburg, 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>.
- [2] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.
- [3] Stephen Boyd. Convex optimization. *Cambridge UP*, 2004.
- [4] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data mining and knowledge discovery*, 21:277–292, 2010.
- [5] Clément L. Canonne. A short note on an inequality between kl and tv, 2023. URL <https://arxiv.org/abs/2202.07198>.
- [6] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffrey Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 339–348, 2019.
- [7] Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.
- [8] Gene H Golub. Some modified matrix eigenvalue problems. *SIAM review*, 15(2):318–334, 1973.
- [9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [10] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.
- [11] 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*, pages 35–50. Springer, 2012.

- 339 [12] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical*  
340 *University of Denmark*, 7(15):510, 2008.
- 341 [13] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. Fast threshold tests for detecting  
342 discrimination. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-*  
343 *First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings*  
344 *of Machine Learning Research*, pages 96–105. PMLR, 09–11 Apr 2018. URL [https://](https://proceedings.mlr.press/v84/pierson18a.html)  
345 [proceedings.mlr.press/v84/pierson18a.html](https://proceedings.mlr.press/v84/pierson18a.html).
- 346 [14] Clayton Scott. Calibrated asymmetric surrogate losses. *Electronic Journal of Statistics*, 6(none):  
347 958 – 992, 2012. doi: 10.1214/12-EJS699. URL <https://doi.org/10.1214/12-EJS699>.
- 348 [15] Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roei Aharoni, Ryan Cotterell, and Ponnur-  
349 rangam Kumaraguru. Representation surgery: Theory and practice of affine steering. *arXiv*  
350 *preprint arXiv:2402.09631*, 2024.
- 351 [16] Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. Beyond single  
352 concept vector: Modeling concept subspace in llms with gaussian distribution. *arXiv preprint*  
353 *arXiv:2410.00153*, 2024.