Fairness for AUC via Feature Augmentation

Hortense Fong, Yale University
Vineet Kumar, Yale University
Anay Mehrotra, Yale University
Nisheeth K. Vishnoi, Yale University

November 2021

Abstract

We study fairness in the context of classification where the performance is measured by the area under the curve (AUC) of the receiver operating characteristic. AUC is commonly used when both Type I (false positive) and Type II (false negative) errors are important. However, the same classifier can have significantly varying AUCs for different protected groups and, in real world applications, it is often desirable to reduce such cross-group differences. We address the problem of how to select additional features to most greatly improve AUC for the disadvantaged group. Our results establish that the unconditional variance of features does not inform us about AUC fairness but class-conditional variance does. Using this connection, we develop a novel approach, fairAUC, based on feature augmentation (adding features) to mitigate bias between identifiable groups. We evaluate fairAUC on synthetic and real-world (COMPAS) datasets and find that it significantly improves AUC for the disadvantaged group relative to benchmarks maximizing overall AUC and minimizing bias between groups.

Contents

1 Introduction

Algorithms form the basis of many important decisions in today's business world and society more generally, with a wide range of applications, including screening applicants for jobs (Liem et al. 2018, De-Arteaga et al. 2019), deciding which individuals might be good candidates for a mortgage (Fuster et al. 2020), and determining which defendants in criminal trials obtain bail (Berk et al. 2018). Many such algorithms have been found to be unfair or discriminatory on the basis of legally and socially salient characteristics like race, gender, and age.

Given the importance of achieving fairness across individuals and groups, a wide range of fair algorithms have been proposed. Most of the fairness interventions assume that data is already collected and fixed, and focus on how to design algorithms that are fair. However, if the original data features are collected without recognizing fairness issues, focusing on only the algorithm might not be sufficient. Consider a scenario in which features are selected to maximize accuracy in a population with two groups. Then the features may be perfectly predictive for the majority group but entirely uninformative for the minority group. A survey of industry practitioners also finds that it is at the data collection step that practitioners seek guidance (Holstein et al. 2019).

Motivated by this problem, we propose a procedure that uses *feature augmentation* (additional feature collection) to improve the predictive performance of disadvantaged groups. Our approach reduces bias, characterized as the difference in the area under the receiver operating characteristic curve (AUC) across the protected groups. Our approach, which we call fairAUC, is applicable to a wide variety of classification algorithms and requires only a few data distribution moments of the additional (auxiliary) features. The method is flexible enough to allow decision-makers or managers to determine where in the fairness-accuracy tradeoff they would like to be.

AUC is a non-parametric performance measure that has long been used in binary classification problems, across a wide range of fields, including diagnostic systems, medicine, and in machine learning (Thompson and Zucchini 1989, Bertsimas et al. 2016, Ahsen et al. 2019). AUC is derived from the receiver operating characteristic (ROC) curve, which captures classifier performance in two dimensions by plotting the true positive rate against the false positive rate, by varying the classification threshold. Integrating the area under the ROC curve summarizes the true positive and false positive rates into a single metric, the AUC. AUC is also related to the Mann-Whitney U statistic, and represents the probability that a *classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance*.

When should a manager use AUC as a model performance criterion? First, classification algorithms require the manager to set a threshold on scores output by a model to separate the classes. AUC provides a threshold-invariant way to obtain model performance without human judgment regarding appropriate thresholds. AUC integrates across thresholds, and is especially useful in environments where there may be multiple managers, who have different thresholds. Second, AUC is invariant to the base rate, or the proportion of individuals in each class. For data with significant class imbalance, an algorithm would achieve high accuracy by simply always predicting the majority class. However, AUC would not assign this algorithm a high performance measure because the algorithm fails to discriminate between the positive and negative classes. Unlike accuracy, AUC is robust to changes in the base rate, which may vary significantly over time and place. Third, AUC serves as a measure of rank-ordering, which is particularly useful when there are different intensities of intervention available (e.g., prescribing medicine vs. performing surgery).

Our Contribution

We propose the fairAUC approach based on feature augmentation to maximally increase the AUC of the disadvantaged group. It allows the manager to identify new (costly) features to acquire. We evaluate the performance of fairAUC alongside benchmark algorithms using synthetic data as well as a real empirical context. We find that fairAUC achieves low bias between groups, while obtaining relatively high levels of AUC. Moreover, our approach permits flexibility in determining how many features to acquire, and suggests which ones, based on AUC, fairness, or a weighted combination.

Feature Augmentation using fairAUC

We use a binormal framework to characterize the distribution of a feature, and show how Fisher's linear discriminant (FLD) can be used to select additional features to reduce bias. FLD produces the linear

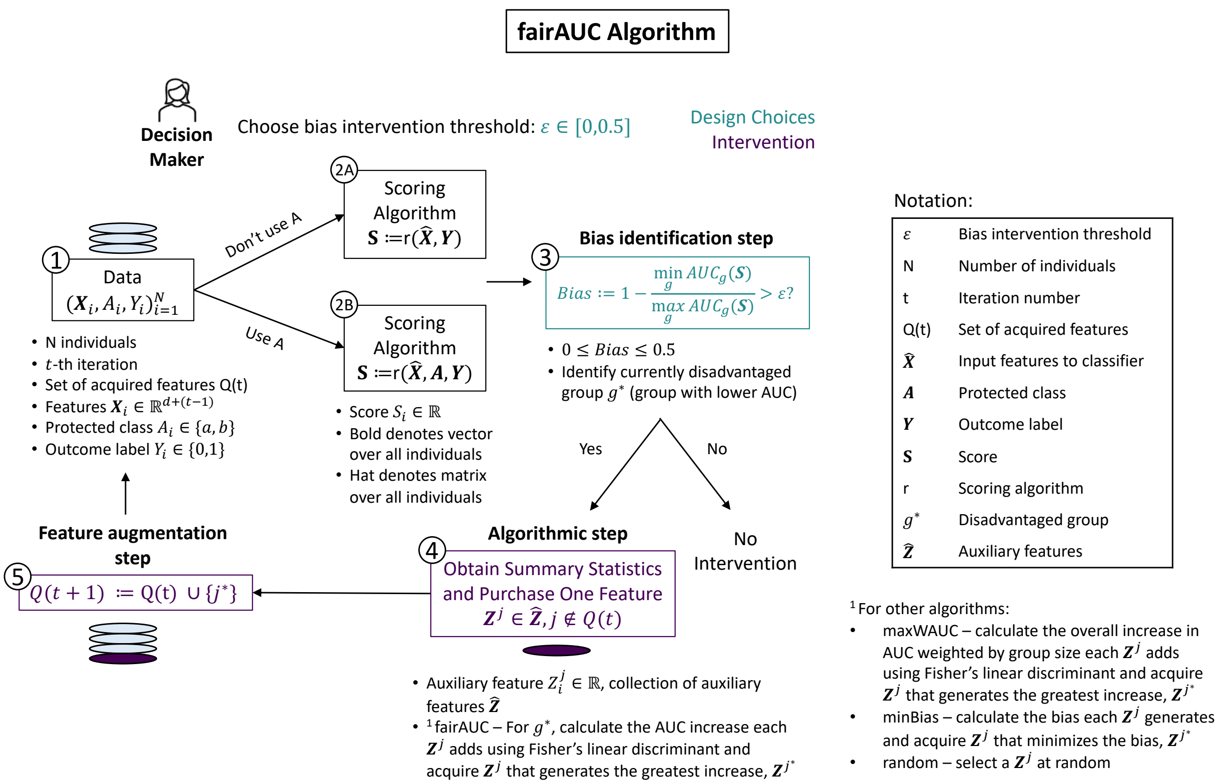


Figure 1: Schematic of our proposed fairAUC feature augmentation procedure.

projection which maximizes AUC within this framework (Su and Liu 1993). While other papers have also suggested searching for additional features to increase fairness (Hardt et al. 2016, Chen et al. 2018), we provide specific recommendations on *which features to select*.

Figure 1 overviews our proposed fairAUC algorithm. fairAUC seeks to improve the AUC of the lower-AUC group, rather than explicitly minimizing bias because the latter does not encourage learning in the long run. Our approach is dynamic, focusing on feature augmentation, whereby we start with a set of initial features and then obtain *additional features* over rounds. The set of features available is summarized by: (a) moments of the data distribution, and (b) correlation with the data already collected. The fairAUC procedure chooses the feature that most increases the AUC of the lower-AUC group, and proceeds through multiple rounds until a threshold condition is satisfied. We also consider three natural benchmarks: minBias, which aims to directly minimize the bias in AUC across groups each round, maxWAUC (max weighted AUC), which ignores fairness constraints to maximize overall AUC across groups, and a random feature selection approach. Each algorithm can be used with or without access to the protected class attribute during classification.

Advantages of fairAUC

The fairAUC procedure has several appealing aspects. First, it can be used with a variety of classification algorithms, and only needs access to a score for each observation, rather than the internal details of the algorithm that produces the score. Second, it uses minimal summary statistics of the augmented features, rather than requiring full access to the feature matrix. One potential source of auxiliary features is from data brokers, which includes a wide array of companies, ranging from White Pages to Acxiom. Third, fairAUC does not treat either of the groups as permanently disadvantaged (or advantaged), unlike most research in the fairness literature. Rather, as we proceed with feature augmentation, the *currently higher-AUC group can become the disadvantaged group after the addition of a new feature*. Thus, our goal in each round is to equalize the AUCs by improving the AUC of the *currently disadvantaged* group, preventing reverse discrimination.

Performance of fairAUC

We first characterize how the fairAUC procedure reduces bias by improving the AUC of the lower-AUC group by a minimum threshold amount, thus providing theoretical performance guarantees. Next, we evaluate the performance of these algorithms with synthetic data generated using a systematic data generation procedure proposed by Guyon (2003). fairAUC achieves significantly greater levels of fairness (in terms of equalizing AUC), with fairly low tradeoffs in AUC. We characterize the accuracy-fairness tradeoff that is achievable using a weighted combination of fairness and AUC objectives, and find that fairAUC obtains low levels of bias without significant sacrifice of overall AUC. Finally, we also evaluate the algorithms using COMPAS recidivism data, a commonly used dataset in fairness studies. We find similar to the synthetic data that fairAUC obtains greater fairness, accompanied by a relatively low tradeoff in accuracy.

2 Related Literature

This paper touches on two streams of literature: fairness in algorithmic systems, a newer and continually growing literature, and AUC of ROC.

2.1 Fairness

The fairness literature addresses questions around bias identification as well as bias reduction.

Sources of Bias

Researchers have documented a number of causes of bias (Barocas and Selbst 2016). They have documented both human (Mejia and Parker 2021) and algorithmic discrimination (Fu et al. 2020). It is critically important to understand the source of bias in order to provide guidance to firms and policymakers on how to address bias, since the recommended intervention would depend on the cause. For example, Lambrecht and Tucker (2019) find that advertising on Facebook with the objective of maximizing cost effectiveness inadvertently shows STEM career ads less frequently to women than men, and they report that the source of this bias is that the market bids up the advertising rates to reach women higher than that for men. Thus, in this case market forces are potentially the cause rather than an algorithm. Our study specifically considers that bias can arise due to the nature of data collected, rather than the algorithm. We focus on feature selection and its impact on classification performance for members of different protected groups.

Fairness Criterion

To quantify bias, a measure relevant to the problem must be used. Several fairness criteria have been proposed. In general, the various measures aim to achieve specific criteria, namely independence (Dwork et al. 2012, Kamiran and Calders 2012, Feldman et al. 2015), sufficiency (Chouldechova 2017), and separation (Hardt et al. 2016, Zafar et al. 2017, Kallus and Zhou 2019). It has been shown under mild assumptions that no measure of fairness can simultaneously achieve two of the three criteria (Kleinberg et al. 2017, Chouldechova 2017, Barocas et al. 2019). Therefore, the appropriate fairness criterion depends on the problem of interest (see Section A for a comparison of measures). Other measures distinguish between individual versus group fairness, and intertemporal ideas of fairness Gupta and Kamble (2019). The criterion we focus on is separation, which recognizes that the protected attribute may be correlated with the target variable. For example, the base rates of loan repayment may differ among groups so a bank may be justified in having different lending rates for different groups (Barocas et al. 2019). The fairness measure we use is related to equalized odds, which achieves separation, in that it is also derived from the ROC curve. Our focus, however, is equalized AUCs, also known as accuracy equity in the literature.

Bias Reduction Strategies

Bias reduction strategies can occur prior to model training by adjusting the feature space (pre-processing), during model training (in-processing), or after model training (post-processing). Pre-processing strategies alter the feature space to be uncorrelated with the protected attribute (Kamiran and Calders 2012, Zemel et al. 2013, Feldman et al. 2015, Celis et al. 2020, Shimao et al. 2019). In-processing strategies directly incorporate the fairness constraint into the optimization problem (Dwork et al. 2012, Zafar et al. 2017, Woodworth et al. 2017, Celis et al. 2019). Using AUC as a fairness metric with in-processing has proven challenging and remains an open problem (Celis et al. 2019). Post-processing strategies occur after classifier training and manipulate the classifier to be uncorrelated with the protected attribute (Hardt et al. 2016). Noriega-Campero et al. (2019) demonstrate that post-processing strategies that rely on randomization to achieve equalized odds are inefficient and Pareto sub-optimal. Most proposed strategies assume the dataset to be fixed and take an algorithmic approach to reducing bias but practitioners have voiced a need for data collection guidance (Holstein et al. 2019). We take a different approach by developing a procedure for additional feature acquisition, which occurs during the data collection stage. Our solution provides guidance on which additional features should be acquired to improve the AUC of the lower-AUC group and ultimately equalize AUCs across groups.

2.2 Area Under the ROC Curve (AUC)

The ROC plots the true positive rate against the false positive rate, visually depicting the tradeoff between the two measures. Integrating the area under the ROC curve produces the AUC, summarizing the ROC into one number (from 0 to 1) with a higher AUC being preferable. AUC also represents the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.

When base rates vary, measures like accuracy, F1, and the area under the precision-recall curve will change even if the fundamental characteristics of the classes remain the same (i.e., $\Pr_{\text{train}}[X|Y] = \Pr_{\text{test}}[X|Y]$ but $\Pr_{\text{train}}[Y] \neq \Pr_{\text{test}}[Y]$). Here, the fact that AUC is insensitive to differences in the base rate of positive instances between the train and test sets is very helpful in a number of settings, including time-varying base rates and cross-sectional (or locational) heterogeneity (Fawcett 2006). Threshold-invariance is useful when the manager seeks to withhold judgement on the classification threshold and when there are multiple managers with varying thresholds.

The probabilistic interpretation of AUC measures the ability of a model to correctly rank order individuals. This is especially valuable when different interventional resource intensities are available (Kallus and Zhou 2019). For example, a radiologist may set different thresholds for different treatment recommendations based on the outcomes of some tests. Similarly, a bank may set different interest rates depending on credit score and other factors. Thus, the manager would be interested in multiple thresholds, not just one, and AUC can provide an overall characterization across all such thresholds.

3 Preliminaries and Assumptions

3.1 Preliminaries

We consider a standard binary classification problem with two groups. The dataset consists of N i.i.d. data points $(X_i, A_i, Y_i)_{i=1}^N$ sampled from a distribution $\mathcal{D}$. Here the input feature $X_i \in \mathbb{R}$, the group $A_i \in \{a, b\}$, and the class label $Y_i \in \{0, 1\}$. Note that here we specify X_i as a scalar "score" for notational simplicity, whereas our fairAUC procedure accommodates general vectors $X_i \in \mathbb{R}^d$.[1] The input feature can represent a single continuous feature or the output of a score function (e.g., logistic regression), which maps multiple input features onto a single real number. Here, and subsequently, we drop the subscript from (X_i, A_i, Y_i) when we do not want to refer to a specific individual. Let $p_{g0}(x) := \Pr_{(X,A,Y)\sim\mathcal{D}}[X = x|A = g, Y = 0]$ denote the distribution of the input feature belonging to the negative class for each group. Similarly, let $p_{g1}(x) := \Pr_{(X,A,Y)\sim\mathcal{D}}[X = x|A = g, Y = 1]$ denote the distribution of the input feature belonging to the positive class for each group.

For a data point (X, A, Y), consider a binary classifier r that, for a threshold $\tau \in \mathbb{R}$ is defined as:

$$r(X, A, Y) := \begin{cases} 1, & \text{if } X \geq \tau \\ 0, & \text{if } X < \tau. \end{cases}$$

[1] For example, with a logistic regression specified as $P(y = 1|x) = \dfrac{\exp(x\theta)}{1 + \exp(x\theta)}$, the score would be $s = x\theta$, and new data that is above a score threshold would be classified as 1.

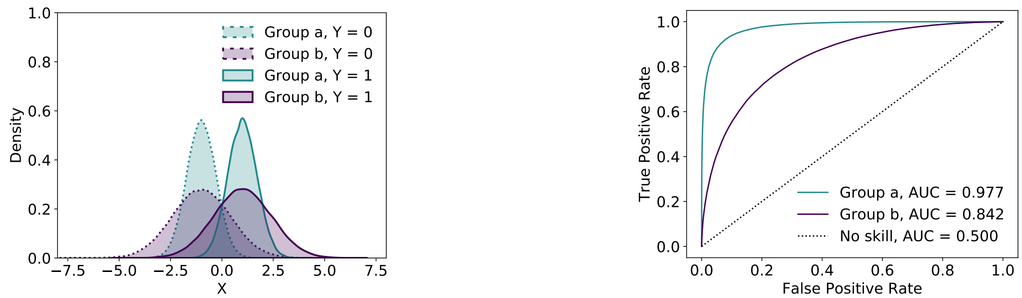


Figure 2: (Left) Binormal density plots where the class means of groups a and b are equal but the class-conditional variances of b are greater than those of a. (Right) ROC curves and AUC by group.

The true positive rate (TPR) measures the proportion of individuals in the positive class being correctly classified as positive. The false positive rate (FPR) measures the proportion of individuals in the negative class being incorrectly classified as positive. Thus, the TPR and FPR are bounded below by 0 and bounded above by 1.[2] These two give rise to the ROC curve as follows: the TPR maps the threshold τ to the y-axis and the FPR maps τ to the x-axis. Formally, for a group g, ROC is defined as $\mathrm{ROC}_g(\tau) \coloneqq (\mathrm{FPR}_g(\tau), \mathrm{TPR}_g(\tau))$. The area under the two-dimensional ROC curve (AUC) aggregates the information captured in the TPR and FPR and is defined for a group g as follows:

Definition 3.1 (Area under the ROC curve (AUC)).

$$\mathrm{AUC}_g \coloneqq \int_0^1 \mathrm{TPR}_g(\mathrm{FPR}_g^{-1}(x))dx. \tag{2}$$

AUC ranges from 0, which occurs when the classifier predicts the opposite of the class label, to 1, which occurs when the classifier can perfectly classify the two classes.

We measure bias by comparing the AUCs obtained from the groups $g \in \{a, b\}$:

Definition 3.2 (Bias).

$$\mathrm{Bias} \coloneqq 1 - \frac{\min_g(\mathrm{AUC}_g)}{\max_g(\mathrm{AUC}_g)}. \tag{3}$$

Bias ranges from 0 to 0.5, with larger values representing greater inequality between groups.

3.2 Class-conditional Means, Variances, and the Binormal Assumption

The class-conditional means and variances of X for each group g are defined as $\mu_{gy} \coloneqq \mathbb{E}[X|A = g, Y = y]$ and $\sigma_{gy}^2 \coloneqq \mathrm{Var}[X|A = g, Y = y]$, respectively. The unconditional (class-independent) variance of X for each group is $\mathrm{Var}[X|A = g]$.

To obtain an analytical relationship between moments of the data and AUC, we assume that for each group the input feature follows a binormal distribution (Pesce and Metz 2007), since it is known to be robust to departures from this assumption (Hanley 1996). This binormal assumption produces four Gaussian distributions: $\mathcal{N}(\mu_{a0}, \sigma_{a0}^2)$, $\mathcal{N}(\mu_{a1}, \sigma_{a1}^2)$, $\mathcal{N}(\mu_{b0}, \sigma_{b0}^2)$, $\mathcal{N}(\mu_{b1}, \sigma_{b1}^2)$. We assume the means of the positive classes are greater than the means of the negative classes for each group ($\mu_{a1} \geq \mu_{a0},\ \mu_{b1} \geq \mu_{b0}$) and that the conditional variances are positive. Figure 2 (Center) displays a density plot of two binormal distributions for which the class-conditional variances within each group are equal but the class-conditional variances for group a are smaller than those for group b.

Incorporating the binormal assumption, TPR and FPR from Equations (1)for group g can be written as:

$$\mathrm{TPR}_g(\tau) = 1 - \Phi\left(\tfrac{\tau - \mu_{g1}}{\sigma_{g1}}\right) \quad \text{and} \quad \mathrm{FPR}_g(\tau) = 1 - \Phi\left(\tfrac{\tau - \mu_{g0}}{\sigma_{g0}}\right) \tag{4}$$

[2]The TPR and FPR of group $g \in \{a, b\}$ can be written as functions of the threshold τ:

$$\mathrm{TPR}_g(\tau) \coloneqq \int_\tau^\infty p_{g1}(x)dx \qquad \text{and} \qquad \mathrm{FPR}_g(\tau) \coloneqq \int_\tau^\infty p_{g0}(x)dx \tag{1}$$

where $\Phi(\cdot)$ represents the standard normal cumulative distribution function. We can now express the AUC defined in Equation (2) as a function of the class-conditional means and variances of each group:

$$\text{AUC}_g = \Phi\left(\frac{\mu_{g1}-\mu_{g0}}{\sqrt{\sigma_{g0}^2+\sigma_{g1}^2}}\right). \tag{5}$$

Figure 2 (Right) displays the ROC curves and their associated AUCs from the binormal distributions shown in Figure 2 (Center). The diagonal line represents random guessing.

4 Methodology

4.1 Unconditional Variance Does not Inform AUC

During exploratory analysis, managers typically analyze the unconditional distributions of the input feature for each group (Corbett-Davies and Goel 2018, Chen et al. 2018, Emelianov et al. 2020). It may be expected that higher variance, flatter unconditional distributions generate higher AUCs since greater spread provides more information. Suppose X takes just one value (is deterministic). Then the unconditional variance is zero and the classifier learns nothing from this data. Given this example, one may believe that a higher unconditional variance corresponds to better classification. However, this thought experiment conflates the separation of means with variance. Analyzing the unconditional distribution mixes together base rates, class-conditional means, and class-conditional variances, obscuring the relationship between the data and AUC.

We formalize the previous ideas by writing the unconditional variance as a function of the conditional variances, where $\pi_g \coloneqq \Pr[Y=1|A=g]$ represents the proportion of observations from the positive class for group g:

$$\text{Var}[X|A=g] \quad = \quad \pi_g(1-\pi_g)(\mu_{g1}-\mu_{g0})^2 + \pi_g\sigma_{g1}^2 + (1-\pi_g)\sigma_{g0}^2. \tag{6}$$

See Section B for the derivation details. When we hold the difference in class means and base rate constant, different combinations of σ_{g0}^2 and σ_{g1}^2 can produce the same unconditional variance in Equation (6). According to Equation (5), the binormal AUC formula, these combinations of σ_{g0}^2 and σ_{g1}^2 do not all map to the same AUC for group g. Indeed, the same unconditional variance can be mapped to multiple AUCs. Table 1 columns 1-4 (Constant $\text{Var}[X|A=g]$) show a numerical example of a single unconditional variance mapping to multiple AUCs for different conditional variances. In the table, $\pi_g = 0.8$ and $\mu_{g1} - \mu_{g0} = 10$. In addition, there is not a monotonic relationship between the unconditional variance and AUC. Columns 5-8 (Increasing $\text{Var}[X|A=g]$) show that increasing the unconditional variance can increase or decrease AUC.

Table 1: Unconditional Variance and AUC Examples

Constant $\text{Var}[X\|A=g]$				Increasing $\text{Var}[X\|A=g]$			
σ_{g0}^2	σ_{g1}^2	Var	AUC	σ_{g0}^2	σ_{g1}^2	Var	AUC
10	1	18.80	0.82	2	4	19.60	0.95
4	2.5	18.80	0.94	12	3	20.80	0.75
2	3	18.80	0.98	4	8	23.20	0.80

Observation 4.1 (Non-informativeness of Unconditional Variance). *The ranking of the unconditional variance between groups is not informative of the ranking of AUC between groups. For groups a and b, if* $\text{Var}[X|A=a] > \text{Var}[X|A=b]$, AUC_a *can be greater than, equal to, or less than* AUC_b.

See Section C for the proof of Observation 4.1.

4.2 Class-conditional Variance Informs AUC

Next, we highlight which features of the data distributions pin down AUC. Equation (5) informs us that increasing the difference in class means and decreasing the class-conditional variances increase the AUC.

Figure 2 (Center) and (Right) visualize an example in which the differences in class means are equal between the two groups but the conditional variances of group b are larger than those of group a. Because of the larger conditional variances, the AUC of b is lower than the AUC of a. Finally, the base rate π_g does not appear in the AUC formula, reinforcing the idea that differences in base rates between the two groups will not contribute to bias with respect to AUC.

4.3 Strategies for Additional Feature Selection

We consider strategies for selecting additional costly features for classification. One natural strategy would be to select features that minimize the bias across groups. We consider a greedy strategy, minBias, that chooses the feature minimizing the difference between AUCs across groups in each round. However, such a strategy fails to incentivize additional learning; when features that are relatively uninformative for both groups minimize bias, it will select those features rather than features that improve AUC.

Another strategy is to select features that improve the AUC of the disadvantaged group (group with lower AUC), decreasing bias in the process. We develop a procedure which does exactly this, noting that in a dynamic feature addition over rounds, the group that is (dis)advantaged may change across rounds. Note that such a procedure does not guarantee a reduction in bias because it is possible that the additional feature is even more predictive for the higher-AUC group.

Thus, the natural question is: given the data we have, what additional features(s) should we acquire to *most improve the AUC of the currently disadvantaged group*? Because AUC is calculated using a scalar score, we must select a dimensionality reduction method which aggregates our existing data with the additional feature(s) into one dimension. We use Fisher's linear discriminant (FLD) as our dimensionality reduction method because it generates the linear projection which maximizes AUC (Su and Liu 1993).

4.3.1 Fisher's Linear Discriminant (FLD)

We make a few simplifying assumptions. First, we assume only *one* additional feature can be acquired in each round and develop a greedy strategy. Second, we assume the existing input feature the manager has can be represented in one dimension (i.e., the output of a score function). Third, we assume that a binormal distribution provides a reasonable approximation for the features. Using FLD, we determine the benefit of a new feature to the AUC of the disadvantaged group under consideration (denoted $g^\star$).

Let n denote the number of individuals in the disadvantaged group $g^\star$. Consider a dataset $(X_i, A_i, Y_i)_{i=1}^n$ where we have already collected one (non-protected) feature $X_i \in \mathbb{R}$ for each individual i. Since our focus is only on a single group, we drop the group subscript notation used in the previous sections. Moreover, assume we have access to an auxiliary feature $(Z_i)_{i=1}^n$, which we could choose to acquire.

We seek to determine the benefit to $g^\star$ of acquiring $(Z_i)_{i=1}^n$. For each outcome class $y \in \{0,1\}$, the class-conditional mean vector and covariance matrix of $(X_i, Z_i)_{i=1}^n$ are:

$$\boldsymbol{\mu_y} := \begin{bmatrix} \mu_{X,y} \\ \mu_{Z,y} \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X|Y=y] \\ \mathbb{E}[Z|Y=y] \end{bmatrix}$$

$$\boldsymbol{\Sigma_y} := \begin{bmatrix} \sigma^2_{X,y} & \rho_y \sigma_{X,y}\sigma_{Z,y} \\ \rho_y \sigma_{X,y}\sigma_{Z,y} & \sigma^2_{Z,y} \end{bmatrix}$$

where $\sigma^2_{X,y} = \mathrm{Var}[X|Y=y]$, $\sigma^2_{Z,y} = \mathrm{Var}[Z|Y=y]$, and $\rho_y = \frac{\mathrm{Cov}[X,Z|Y=y]}{\sigma_{X,y}\sigma_{Z,y}}$. Note that ρ_y represents the class-conditional correlation of X and Z and not the unconditional correlation.

Let $\boldsymbol{w}$ represent a potential projection direction that projects (X_i, Z_i) for each individual $i \in [n]$ to $\mathbb{R}$, combining the two features into a single value. Then the projected class mean $\tilde{\mu}_y \in \mathbb{R}$ is defined as $\tilde{\mu}_y := \boldsymbol{w}^\top \boldsymbol{\mu_y}$ and the projected class-conditional variance $\tilde{\sigma}_y^2 \in \mathbb{R}$ is defined as $\tilde{\sigma}_y^2 := \boldsymbol{w}^\top \boldsymbol{\Sigma_y} \boldsymbol{w}$. The FLD objective function is known to maximize AUC (Su and Liu 1993). In terms of the projected means and variances, the FLD objective is: $J(\boldsymbol{w}) := \frac{(\tilde{\mu}_1 - \tilde{\mu}_0)^2}{\tilde{\sigma}_0^2 + \tilde{\sigma}_1^2}$. In terms of the pre-projection mean vectors and covariance matrices, it is: $J(\boldsymbol{w}) = \frac{\boldsymbol{w}^\top(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^\top \boldsymbol{w}}{\boldsymbol{w}^\top(\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})\boldsymbol{w}}$. The projection direction $\boldsymbol{w}$ which maximizes $J(\boldsymbol{w})$ can be found by solving a generalized eigenvalue problem (Duda et al. 2012). The optimal linear projection direction (when $\mathbf{\Sigma_0} + \mathbf{\Sigma_1}$ is invertible) is given by:

$$\boldsymbol{w}^\star = (\mathbf{\Sigma_0} + \mathbf{\Sigma_1})^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0}). \tag{7}$$

Plugging $\tilde{\mu}_y = \boldsymbol{w}^{\star\top}\boldsymbol{\mu_y}$ and $\tilde{\sigma}_y^2 = \boldsymbol{w}^{\star\top}\mathbf{\Sigma_y}\boldsymbol{w}^\star$ into Equation (5) yields the AUC of the optimal linear combination of input features (X, Z):

$$\mathrm{AUC}(X, Z) = \Phi\left(\sqrt{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^\top(\mathbf{\Sigma_0} + \mathbf{\Sigma_1})^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})}\right). \tag{8}$$

The benefit to the disadvantaged group of acquiring Z is the difference between this new value of AUC and the previous value of AUC that used only X.

So far, Z has represented an arbitrary feature available for acquisition. When given a choice over many possible features, which feature Z maximizes Equation (8) for a given X?

4.4 fairAUC Procedure

We now present our fairAUC procedure, which is a greedy procedure which helps managers determine which additional features should be acquired to maximally increase the AUC of the lower AUC group. Equation (8) serves as the backbone for our proposed fairAUC procedure. It relies on knowing only a few summary statistics of the data. It can be used with data sellers who provide costly features. Alternatively, managers may collect a small sample of additional features and estimate the benefit of each using this strategy prior to collecting the features for all individuals.

It begins by taking in the data the manager has for N individuals $(\boldsymbol{X}_i, A_i, Y_i)_{i=1}^N$, the manager's scoring algorithm r, and the level of acceptable bias ε. As before, we drop the subscript from $(\boldsymbol{X}_i, A_i, Y_i)$ when we do not refer to a specific individual. The input data $\boldsymbol{X} \in \mathbb{R}^d$, the group $A \in \{a, b\}$, and the class $Y \in \{0, 1\}$.

The manager aggregates the features in $\boldsymbol{X}$ into a single score $S \in \mathbb{R}$ using a fixed scoring algorithm r. The scoring algorithm that most aligns with our fairAUC procedure is FLD but our framework allows for any scoring algorithm r. Moreover, r may or may not use the protected attribute A depending on the context. For instance, FICO is prohibited from using characteristics like race, gender, and marital status in producing its credit score. We refer to using A as using separate classifiers for each group and not using A as using only a single classifier for both groups.

In each round t of the fairAUC procedure, there is a bias identification step, an algorithmic step, and a feature augmentation step. In the bias identification step, we first calculate the AUC for each of the groups from the scores S. The bias of the model is calculated from the AUCs of the two groups. If the bias is smaller than a given tolerance level ε, the manager does not need to take any intervention to reduce bias. However, if the bias is larger than ε, the manager acquires one additional feature per round. The fairAUC procedure follows a greedy approach. The group with lower AUC is referred to as the currently disadvantaged group and is denoted $g^\star$. The group considered disadvantaged can vary over the rounds of feature acquisition.

In the algorithmic step, fairAUC aims to acquire the feature that most increases the AUC of $g^\star$ using the FLD heuristic explained in the previous section (see Equation (8)). As previously discussed, FLD generates a linear combination of features which maximizes AUC and requires only summary statistics to calculate. Let $(\boldsymbol{Z_i})_{i=1}^N$ where $\boldsymbol{Z_i} \in \mathbb{R}^{d'}$ represent the auxiliary features available for acquisition. Let $m = d + d'$ capture the total number of features that exist in the data the manager owns and in the auxiliary data available for acquisition. Let $\{1, \ldots, d'\}$ denotes the indices of all auxiliary features that are available for acquisition. In any given round t, we use $Q(t) \subset \{1, \ldots, d'\}$ to denote the set of auxiliary features acquired features so far. Initially, we have that $Q(0) = \emptyset$. Hence, the set of features available is $[d'] \setminus Q$. We assume that the cost of each feature is the same and that a feature is acquired for all N individuals.

For the group $g^\star$, the manager obtains the class-conditional means of each of the features available for acquisition in $[d'] \setminus Q$ as well as the class-conditional covariance matrices of each of the available features and the score S.[3] The conditional means and covariances inform how valuable each of the additional features is

[3]Note that in the case of working with a data seller the fairAUC procedure assumes that the data seller also knows A and Y. In practice, this information may need to be shared. A benefit of the fairAUC procedure is that it does not require the manager to share X with the data seller, only S. It also does not require the data seller to share more than a few summary statistics.

Procedure 1: fairAUC (t-th iteration)

Input: data owned $(\boldsymbol{X_i}, A_i, Y_i)_{i=1}^N$, scoring algorithm r, bias threshold ε, set of acquired features $Q(t)$, data available for acquisition $(\boldsymbol{Z_i})_{i=1}^N$;
Output: $Q(t+1)$;
if $\boldsymbol{A}$ *cannot be used* **then**
    $\boldsymbol{S} \coloneqq r(\hat{\boldsymbol{X}}, \boldsymbol{Y})$;
else
    $\boldsymbol{S} \coloneqq r(\hat{\boldsymbol{X}}, \boldsymbol{A}, \boldsymbol{Y})$;
for *group* $g \in \{a, b\}$ **do**
    compute $\mathrm{AUC}_g(S)$ (Definition 3.1) ;
$g^\star \coloneqq \arg\min_g \mathrm{AUC}_g(S)$ (Disadvantaged group) ;
Bias $\coloneqq 1 - \frac{\min_g(\mathrm{AUC}_g(S))}{\max_g(\mathrm{AUC}_g(S))}$ (Definition 3.2) ;
if Bias $> \varepsilon$ **then**
    for *feature* $\boldsymbol{Z}^j \in \hat{\boldsymbol{Z}}, j \notin Q(t)$ **do**
        for feature $\boldsymbol{Z}^j$, group $g^\star$, and score $\boldsymbol{S}$, obtain class-conditional means, $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$, and covariance matrices, $\boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$ (Summary Statistics by Group Subroutine);
        $h(\boldsymbol{S}, \boldsymbol{Z}^j) \coloneqq \Phi\left(\sqrt{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^\top (\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})}\right)$;
    $j^\star \coloneqq \arg\max_j h(\boldsymbol{S}, \boldsymbol{Z}^j)$;
    acquire feature $\boldsymbol{Z}^{j^\star}$;
    return $Q(t+1) \coloneqq Q(t) \cup \{j^\star\}$;
else
    no intervention;

to the manager in terms of increasing the AUC of $g^\star$. The manager acquires the feature which maximizes the AUC of group $g^\star$.

In the feature augmentation step, $Q(t)$ is updated to include this new feature. The feature acquired is concatenated with the existing dataset and becomes the input to the next round. Procedure 1 formalizes each iteration of fairAUC. To simplify the notation, let $\hat{\boldsymbol{X}}$ denote the collection of $(\boldsymbol{X}_i)_{i=1}^N$, and similarly $\hat{\boldsymbol{Z}}$ the collection of $(\boldsymbol{Z}_i)_{i=1}^N$. $\boldsymbol{Z}_i$ denotes a row vector of features for each individual, $\boldsymbol{Z}^j$ denotes a column feature vector across all individuals, and $\hat{\boldsymbol{Z}}$ denotes the collection of features for all individuals.

Subroutine: Summary Statistics by Group

Input: feature available for acquisition $\boldsymbol{Z}$, group g, existing score $\boldsymbol{S}$;
Output: class-conditional mean vectors $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$, class-conditional covariance matrices $\boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$;
for *class* $y \in \{0, 1\}$ **do**
    $n \coloneqq n_{A=g, Y=y}$;
    $\boldsymbol{\mu_y} \coloneqq \begin{bmatrix} \bar{S}_y \\ \bar{Z}_y \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\sum_{i:\, \substack{A_i=g \\ Y_i=y}} S_i \\ \frac{1}{n}\sum_{i:\, \substack{A_i=g \\ Y_i=y}} Z_i \end{bmatrix}$;
    $\boldsymbol{\Sigma_y} \coloneqq \begin{bmatrix} \sigma^2_{S,y} & \rho_y \sigma_{S,y}\sigma_{Z,y} \\ \rho_y \sigma_{S,y}\sigma_{Z,y} & \sigma^2_{Z,y} \end{bmatrix}$
    where $\sigma^2_{S,y} = \frac{1}{n-1}\sum_{i:\, \substack{A_i=g \\ Y_i=y}} (S_i - \bar{S}_y)^2$, $\sigma^2_{Z,y} = \frac{1}{n-1}\sum_{i:\, \substack{A_i=g \\ Y_i=y}} (Z_i - \bar{Z}_y)^2$, and
    $\rho_y = \frac{1}{(n-1)\sigma_{s,y}\sigma_{Z,y}} \sum_{i:\, \substack{A_i=g \\ Y_i=y}} (S_i - \bar{S}_y)(Z_i - \bar{Z}_y)$;
return $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$

4.5 Theoretical Guarantees on Improvement of AUC by fairAUC

In this section, we present our theoretical result on the fairAUC procedure in the binormal framework. At a high level, our result is based on the idea that we can provide fairness guarantees for fairAUC, which uses

FLD-based scores S (introduced in Section 4.3.1 and defined in Equation (15) in Section F). In each iteration $t \in \mathbb{N}$ of fairAUC, where the AUC for disadvantaged group is bounded away from 1, and there is at least one auxiliary feature present which has "low" class-conditional covariances with the current scores S on the disadvantaged group and has "bounded" class-conditional variances and means on the disadvantaged group, fairAUC is guaranteed to improve the AUC of the disadvantaged group by at least a constant in iteration t.

Proposition 4.2 (Theoretical guarantee on fairAUC; informal statement. See Theorem F.4 for formal version.). *In the binormal framework, if the summary statistics subroutine outputs the unbiased means and covariances of the queried features, then we can provably guarantee that in each iteration of the fairAUC procedure, the AUC value of the disadvantaged group increases by at least* $\max_\ell \frac{1}{18} \cdot \gamma^2 \cdot \beta_\ell^2 \cdot (1 - \delta_\ell)^2$*. Here* ℓ *runs over unacquired features,* γ *is the distance of current AUC value of the disadvantaged group from* 1*,* β_ℓ *is the difference in the normalized class-conditional mean of the* ℓ*th unacquired feature on the disadvantaged group (Equation* (17)*), and* δ_ℓ *is the absolute value of the normalized class-conditional covariance between the score* $\mathbf{S}$ *and the* ℓ*th unacquired feature on the disadvantaged group (Equation* (18)*).*

The theorem implies that the improvement is more when γ and β_ℓ are large and δ_ℓ is close to 0 for at least one unacquired feature, and less when γ is close to 0 or for all features either β_ℓ is close to 0 or δ_ℓ is close to 1. To see why the improvement value may depend on these quantities, note that:

1. If γ is close to 0, then the AUC of disadvantaged group is close to 1, which is its maximum value, and hence, cannot increase significantly. In contrast, when we have data with features that are not as informative, then the potential improvement γ and actual improvement are higher.

2. Intuitively, the difference between class-conditional means (β_ℓ) helps create separation between the two classes, and including features with high separation improves accuracy and AUC. If β_ℓ is close to 0, then Equation (5) tells us that the classifier using the ℓth unacquired feature to predict the outcome has a low AUC, and so the ℓth unacquired feature is not a "good predictor" and does not increase the AUC significantly.

3. The normalized covariance provides a measure of dependence between the new feature to be acquired and the score that summarizes existing features. This is related to mutual information, and since FLD is a linear discriminant, covariance provides a characterization of the mutual information. When δ_ℓ is close to 1, then the ℓth unacquired feature is highly correlated with the score derived from the existing features, and so, does not "add additional information."

There are a few points to note. First, while a specific feature ℓ might create separation for the disadvantaged group, it might also create such separation for the advantaged group as well. Recall that fairAUC by design only focuses on improving the performance of the disadvantaged group. However, we also prove that, fairAUC does not decrease the AUC of the advantaged group. In Theorem F.5, we prove similar bounds for the advantaged group. Second, while fairAUC is guaranteed to pick the "best" feature for the disadvantaged group, it may not pick the best feature for the minority group if the minority group is currently the advantaged one. Finally, note that it is possible for the AUC of the currently disadvantaged group to exceed that of the advantaged group due to this feature acquisition. If that happens, the definition of disavantaged group changes for the next round.

5 Empirical Results

5.1 Synthetic Data

We use the data generation strategy proposed by Guyon (2003) for the controlled benchmarking of variable selection algorithms in binary classification problems. Please see details in Section E.1.

5.1.1 Procedure Comparison

We compare four feature selection strategies, namely fairAUC, maxWAUC, minBias, and random. During each round of feature acquisition, the fairAUC procedure selects the feature that most improves AUC for the

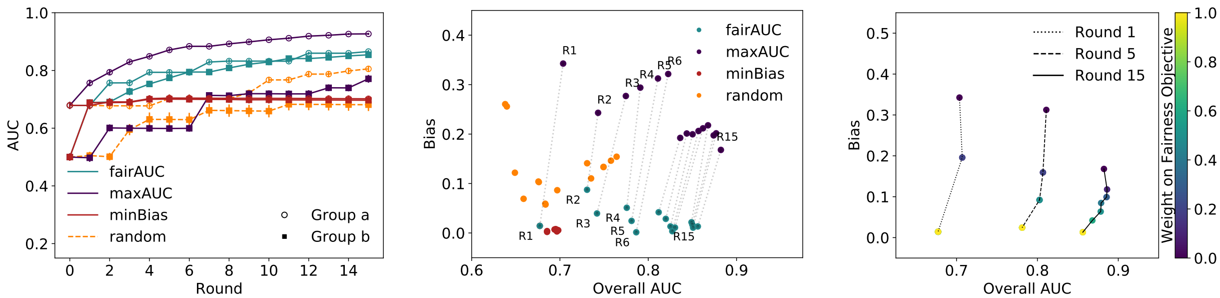


Figure 3: (Left) AUC by group over feature augmentation rounds using different feature acquisition strategies and using the protected attribute. (Center) Comparison of accuracy-fairness tradeoff among feature acquisition strategies using the protected attribute. (Right) Pareto frontier of convex combinations of the fairness and AUC objectives for several rounds of feature augmentation using the protected attribute.

group with lower AUC according to FLD. The maxWAUC procedure selects the feature that most improves the *overall weighted AUC* using FLD weighted by group size (see Section D for the maxWAUC algorithm). The minBias procedure selects the feature that minimizes the bias between the two groups. The random procedure selects a feature at random, and represents a baseline in which the manager collects additional data in an uninformed manner.

5.1.2 Synthetic Data Results

Figure 3 (Left) compares the fairAUC, maxWAUC, minBias, and random procedures when the protected attribute is used in the scoring function. This is equivalent to each group having a separate classifier. Under fairAUC, the initially disadvantaged minority group b quickly obtains predictive performance equal to group a. Under maxWAUC, group b's AUC always trails group a's AUC even though separate classifiers are trained for each group. The minBias procedure quickly reduces bias and maintains low bias but fails to select informative features. fairAUC and maxWAUC outperform the random procedure. Figure 5 in Section E compares the procedures when the protected attribute is not used. The overall patterns among the procedures remain the same. Compared to Figure 3 (Left), the achieved AUCs from fairAUC and maxWAUC are lower because of the shared weights between the two groups.

Accuracy is known to monotonically increase with AUC (Cortes and Mohri 2003). We graph the accuracy-fairness tradeoff (where accuracy is measured by AUC) in Figure 3 (Center) that results from using fairAUC rather than maxWAUC. Ideally, a procedure generates points in the lower right of the graph, i.e. low bias and high AUC (accuracy). The dotted lines connect the corresponding rounds between the two procedures. All of the lines have a positive slope, indicating that fairAUC reduces bias but at the cost of overall AUC. For fairAUC, we observe that the bias does not monotonically decrease but rather jumps around. After all the rounds are complete, the manager can evaluate the accuracy (AUC) versus bias tradeoff, according to their requirements. If the manager requires a lower bias, they could choose the round that corresponds to the lowest level of bias (Round 6), whereas if they prefer to tradeoff a higher level of bias for a higher AUC, they might choose Round 15. The crucial aspect is that the feature augmentation algorithm provides the manager with a flexible set of options at various points on the accuracy-bias spectrum.

The minBias procedure as expected produces low bias values but at the cost of significantly lower AUC. The random procedure generates bias values between fairAUC and maxWAUC but at far worse AUC values than either. Figure 6 in Section E graphs the tradeoff when the protected attribute A is not used. The lines in Figure 3 (Center) are closer to being vertical, indicating there is less of a tradeoff when the protected attribute can be used.

We evaluate convex combinations of the fairness and maximum AUC objectives to generate a Pareto frontier for bias and overall AUC. Figure 3 (Right) shows the intermediate bias and overall AUC values that can be achieved by altering the weight of the two objectives over different feature augmentation rounds when A is used.[4] Full weight on the fairness objective represents fairAUC and full weight on the AUC objective

[4]For earlier rounds, many weight combinations select the same feature acquisition strategy, resulting in overlap.

represents maxWAUC. The manager therefore also has flexibility in determining how much weight to give to each of the objective functions. Figure 7 in Section E graphs the Pareto frontier when A is not used and shows a similar pattern.

The results highlight a number of pitfalls that can occur in data collection and prediction algorithm design. First, collecting data to maximize overall AUC or accuracy can inadvertently hurt the minority group. This can occur even when the two groups are equally separable and separate classifiers are trained for each group. The selected features under maxWAUC perform similarly to a random selection strategy for the minority group. Second, in practice managerstend to incorporate all data available (Holstein et al. 2019). If a single classifier is trained and additional data disproportionately represents the majority group, the weights will be influenced more by the majority group. Third, a strategy that aims to only minimize bias can result in the collection of features that are not predictive for either group, and can even result in lower accuracy.

5.2 Application: Predicting Violent Recidivism

We use the COMPAS recidivism dataset produced by ProPublica to demonstrate the previous ideas in a real world setting where prediction of future criminal activity can impact the conditions of confinement (Larson et al. 2016).[5]

5.2.1 COMPAS Dataset

The dataset covers over 6,000 criminal defendants from Broward County, Florida and contains information on their COMPAS score, demographics (gender, race, age), criminal history, and whether they actually recidivated within a two-year period after release. Our target variable of interest is violent recidivism and the protected attribute is age (under 25 vs. 25+). Those under 25 represent 33% of the data and have a 14% violent recidivism rate while those over 25 have a 10% violent recidivism rate.

5.2.2 Data Pre-processing

We take log of the numerical variables in the dataset (e.g., number of priors) to reduce the impact of outliers. We also convert the categorical variables (e.g., race) into binary variables. We do not use the risk assessment levels or decile scores generated by COMPAS as inputs since they are the outcome variables, i.e. essentially what we seek to predict.

Suppose a judge has data on gender (initial independent variable), age group (sensitive attribute), and whether each defendant reoffended within two years of being released (outcome variable). Defendants under 25 years of age are the initially disadvantaged group based on the data the judge has. Given that features are costly to acquire, our focus is on which additional feature a judge should collect to better predict the likelihood of recidivism for defendants under 25.

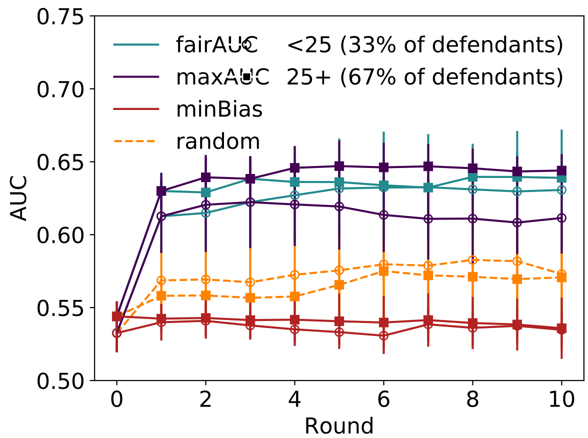
5.2.3 Results

Figure 4 plots the performance of the four procedures for the COMPAS recidivism dataset when age is used. The AUCs have fairly large error bars but the means follow the pattern seen in the synthetic data. fairAUC improves the AUC for the group of defendants under 25 and decreases bias while maxWAUC does not close the gap between the two groups.

Table 2 shows the variables selected by fairAUC and the associated parameters discussed in Theorem F.4. Recall that the parameters that define the lower bound in Theorem F.4 are the measure of separation β_ℓ, the class-conditional covariance with the score δ_ℓ, and the room for improvement γ. At round 8, the disadvantaged group flips before flipping back at round 9. The variables that most greatly improve the AUC of the disadvantaged group (those chosen early) are those with large separations β_ℓ. The benefit from having a larger separation appears to dominate the cost of higher class-conditional covariance with the score δ_ℓ. In general, γ decreases with increasing rounds while the AUC increases. $1-\gamma$ measures the AUC prior to new

[5]With an existing dataset, we are limited to the features collected in the dataset. If the features were collected without fairness in mind, it will be challenging to discover features within the dataset more predictive for the disadvantaged group so the empirical application should be considered a conservative example.

Figure 4: Predicting violent recidivism using protected attribute (age)



feature acquisition while the FLD AUC is based on the FLD approximation after feature acquisition. The difference between $1-\gamma$ and the FLD AUC of the previous round is likely due to the fact that most of the variable are non-continuous indicator variables. Although the majority of the variables are not continuous but instead indicators, the implications of the theorem still hold.

Table 2: COMPAS Features and Theoretical Parameters

Round	Variable ℓ	β_ℓ	δ_ℓ	γ	FLD AUC
1	# Priors	0.2979	0.1935	0.4663	0.6790
2	# Juv. Misdemeanors	0.2222	1	0.3841	0.6991
3	I(Battery)	0.0700	1	0.3773	0.7098
4	I(Possession Meth)	0.1144	0.5948	0.3718	0.7101
5	I(Possession Cannabis)	0.1072	0.0407	0.3667	0.7147
6	I(Burglary)	0.0956	0.0314	0.3607	0.7185
7	I(Hispanic)	0.1106	0.3994	0.3627	0.7162
8	I(Driving License Revoked)	0.1012	0.7279	0.3617	0.7056
9	I(White)	0.0974	0.4000	0.3589	0.7149
10	I(Possession Cannabis Sell)	0.0733	0.1716	0.3537	0.7189

6 Conclusion

We propose fairAUC, an approach to feature augmentation that helps achieve fairness in the AUC measure, which has received little attention from a fairness perspective. Our approach, which can incorporate a wide variety of classification algorithms, aims to improve the performance of each group, in addition to minimizing bias. We demonstrate the value of our method in multiple ways. First, using a theoretical analysis we show provable AUC improvements for the disadvantaged group. Second, we test our approach using synthetic data as well as in a real-world context and find that our approach performs well in reducing bias, while also increasing AUC for both the disadvantaged and advantaged groups.

While our method has many advantages, it is not without limitations. First, our method applies to cases with binary groups and binary outcome labels, although in principle it could be extended to more than two groups and a multiclass classification problem. Second, if two ROC curves cross then one classifier performs better in one region of ROC space and the other classifier performs better in the other region of ROC space. Our approach would only consider the overall AUC. Third, the underlying AUC metric might not be the best for all practical situations, since it weighs both false positives and true positives, whereas one of these might be more important. In practice, the algorithm can be altered to account for such asymmetric weights. Fourth, our procedure assumes the underlying data distributions are approximately binormal. While the fairAUC procedure is meant to provide guidance as a heuristic, large deviations from normality may undermine its effectiveness.

We trust that this paper is a first step in identifying and directly addressing fairness as it relates to the data collection process and AUC, and expect that more broadly these aspects will be further investigated in future research.

References

Mehmet Eren Ahsen, Mehmet Ulvi Saygi Ayvaci, and Srinivasan Raghunathan. When Algorithmic Predictions Use Human-Generated Data: A Bias-Aware Classification Algorithm for Breast Cancer Diagnosis. *Information Systems Research*, 30(1):97–116, 2019.

Solon Barocas and Andrew D Selbst. Big Data's Disparate Impact. *Calif. L. Rev.*, 104:671, 2016.

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

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, page 0049124118782533, 2018.

Dimitris Bertsimas, Allison O'Hair, Stephen Relyea, and John Silberholz. An Analytics Approach to Designing Combination Chemotherapy Regimens for Cancer. *Management Science*, 62(5):1511–1531, 2016.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 319–328, 2019.

L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data Preprocessing to Mitigate Bias: A Maximum Entropy Based Approach. In *International Conference on Machine Learning*, pages 1349–1359. PMLR, 2020.

Irene Y. Chen, Fredrik D. Johansson, and David A. Sontag. Why Is My Classifier Discriminatory? In *NeurIPS*, pages 3543–3554, 2018.

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

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

Corinna Cortes and Mehryar Mohri. AUC Optimization vs. Error Rate Minimization. In *NIPS*, pages 313–320. MIT Press, 2003.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. Wiley, 2012. ISBN 9781118586006. URL `https://books.google.com/books?id=Br33IRC3PkQC`.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness Through Awareness. In *ITCS*, pages 214–226, New York, NY, USA, 2012. ACM.

Vitalii Emelianov, Nicolas Gast, Krishna P Gummadi, and Patrick Loiseau. On Fair Selection in the Presence of Implicit Variance. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 649–675, 2020.

Tom Fawcett. An Introduction to ROC Analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *KDD*, pages 259–268. ACM, ACM, 2015.

Runshan Fu, Yan Huang, and Param Vir Singh. Artificial Intelligence and Algorithmic Bias: Source, Detection, Mitigation, and Implications. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pages 39–63. INFORMS, 2020.

Andreas Fuster, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther. Predictably Unequal? The Effects of Machine Learning on Credit Markets. *The Effects of Machine Learning on Credit Markets (October 1, 2020)*, 2020.

Swati Gupta and Vijay Kamble. Individual Fairness in Hindsight. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 805–806, 2019.

Isabelle Guyon. Design of Experiments of the NIPS 2003 Variable Selection Benchmark. In *NIPS 2003 workshop on feature extraction and feature selection*, volume 253, 2003.

James A Hanley. The Use of the 'Binormal' Model for Parametric ROC Analysis of Quantitative Diagnostic Tests. *Statistics in medicine*, 15(14):1575–1585, 1996.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of Opportunity in Supervised Learning. In *NIPS*, pages 3315–3323, 2016.

Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2019.

Nathan Kallus and Angela Zhou. The Fairness of Risk Scores beyond Classification: Bipartite Ranking and the xAUC Metric. In *Advances in Neural Information Processing Systems*, pages 3438–3448, 2019.

Faisal Kamiran and Toon Calders. Data Preprocessing Techniques for Classification without Discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *ITCS*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

Anja Lambrecht and Catherine Tucker. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of Stem Career Ads. *Management Science*, 65(7):2966–2981, 2019.

Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica (5 2016)*, 9, 2016.

Cynthia CS Liem, Markus Langer, Andrew Demetriou, Annemarie MF Hiemstra, Achmadnoer Sukma Wicaksana, Marise Ph Born, and Cornelius J König. Psychology Meets Machine Learning: Interdisciplinary Perspectives on Algorithmic Job Candidate Screening. In *Explainable and interpretable models in computer vision and machine learning*, pages 197–253. Springer, 2018.

Jorge Mejia and Chris Parker. When Transparency Fails: Bias and Financial Incentives in Ridesharing Platforms. *Management Science*, 67(1):166–184, 2021.

Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle, and Alex 'Sandy' Pentland. Active Fairness in Algorithmic Decision Making. In *AIES*, pages 77–83. ACM, 2019.

Lorenzo L Pesce and Charles E Metz. Reliable and Computationally Efficient Maximum-Likelihood Estimation of "Proper" Binormal Roc Curves. *Academic radiology*, 14(7):814–829, 2007.

Sebastien Roch. Essentials of Modern Discrete Probability: A Toolkit for the Discrete Probabilist. 2014. `https://www.math.wisc.edu/~roch/teaching_files/833.f14/`.

Hajime Shimao, Junpei Komiyama, Warut Khern-am nuai, and Karthik Natarajan Kannan. Strategic Best-Response Fairness in Fair Machine Learning Algorithms. *Available at SSRN 3389631*, 2019.

David Stirzaker. *Elementary Probability*. Cambridge University Press, 2003. ISBN 9781139441032. URL `https://books.google.com/books?id=_HUG6KNwpI4C`.

John Q Su and Jun S Liu. Linear Combinations of Multiple Diagnostic Markers. *Journal of the American Statistical Association*, 88(424):1350–1355, 1993.

Mary Lou Thompson and Walter Zucchini. On the Statistical Analysis of ROC Curves. *Statistics in Medicine*, 8(10): 1277–1290, 1989.

Joel A. Tropp. An Introduction to Matrix Concentration Inequalities. *Found. Trends Mach. Learn.*, 8(1-2):1–230, 2015.

Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning Non-Discriminatory Predictors. In *COLT*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR, 2017.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–325–III–333. JMLR.org, 2013. URL `http://dl.acm.org/citation.cfm?id=3042817.3042973`.

A Measures of Fairness Used in the Literature

Table 3 provides advantages and disadvantages to several fairness metrics that have been suggested in the literature. Let $Y \in \{0,1\}$ represent the true outcome, $\hat{Y} \in \{0,1\}$ the predicted outcome, A the protected attribute, X the non-protected attributes, and C the classifier.

Table 3: Measures of Fairness in the Literature

Measure	Definition, Advantages, Disadvantages
Unawareness	$C = C(X)$ Advantage: Addresses disparate treatment and complies with existing laws (e.g., Civil Rights Act of 1964) by not using race as an explicit variable. Disadvantage: If X and A are correlated then the protected attribute is still incorporated into the classifier.
Statistical Parity	$P(\hat{Y}=1\|A=i) = P(\hat{Y}=1\|A=j)$ $\forall$ groups i, j Advantage: Addresses disparate impact and is the foundation for some laws (e.g., four-fifths rule). Disadvantages: Can be achieved simply by selecting x% from all groups regardless of justification, potentially resulting in reverse-discrimination. If Y and A are correlated, the ideal predictor $\hat{Y} = Y$ cannot be obtained.
Predictive Rate Parity	$P(Y=1\|A=i,\hat{Y}=1) = P(Y=1\|A=j,\hat{Y}=1)$ $\forall$ groups i, j and $P(Y=0\|A=i,\hat{Y}=0) = P(Y=0\|A=j,\hat{Y}=0)$ $\forall$ groups i, j Advantage: Optimality-compatible (i.e., allows $\hat{Y} = Y$), aligning fairness with accuracy, and avoids reverse-discrimination. Disadvantage: May not close the gap between groups over time if Y and A are correlated.
Equalized Odds	$P(\hat{Y}=1\|A=i,Y=1) = P(\hat{Y}=1\|A=j,Y=1)$ $\forall$ groups i, j and $P(\hat{Y}=1\|A=i,Y=0) = P(\hat{Y}=1\|A=j,Y=0)$ $\forall$ groups i, j Advantage: Optimality-compatible (i.e., allows $\hat{Y} = Y$) and avoids reverse-discrimination. Disadvantage: May not close the gap between groups over time if Y and A are correlated.

B Unconditional Variance as a Function of the Conditional Variances

Let the proportion of observations from the positive class for group g be

$$\pi_g := \Pr[Y = 1|A = g].$$

The conditional variances, σ_{g1}^2 and σ_{g0}^2, and the unconditional variance, $\mathrm{Var}[X|A = g]$, can be written as:

$$\begin{aligned} \sigma_{g1}^2 &= \mathrm{Var}[X|A = g, Y = 1] \\ &= \mathbb{E}[X^2|A = g, Y = 1] - \mathbb{E}[X|A = g, Y = 1]^2 \\ &= \int x^2 p_{g1}(x)dx - \mu_{g1}^2, \end{aligned} \tag{9}$$

$$\begin{aligned}\sigma_{g0}^2 &= \mathrm{Var}[X|A=g, Y=0] \\ &= \mathbb{E}[X^2|A=g, Y=0] - \mathbb{E}[X|A=g, Y=0]^2 \\ &= \int x^2 p_{g0}(x)dx - \mu_{g0}^2, \end{aligned} \tag{10}$$

and

$$\begin{aligned}\mathrm{Var}[X|A=g] &= \mathbb{E}[X^2|A=g] - \mathbb{E}[X|A=g]^2 \\ &= \pi_g \int x^2 p_{g1}(x)dx + (1-\pi_g)\int x^2 p_{g0}(x)dx - (\pi_g \mu_{g1} + (1-\pi_g)\mu_{g0})^2.\end{aligned} \tag{11}$$

It follows from Equations (9), (10), and (11) that:

$$\mathrm{Var}[X|A=g] \quad = \quad \pi_g(1-\pi_g)(\mu_{g1}-\mu_{g0})^2 + \pi_g\sigma_{g1}^2 + (1-\pi_g)\sigma_{g0}^2.$$

C Proof of Observation 4.1

Proof of Observation 4.1. Since $\pi_a = \pi_b = \pi$, $\mu_{a1} = \mu_{b1} = \mu_1$, and $\mu_{a0} = \mu_{b0} = \mu_0$,

$$\mathrm{Var}[X|A=a] > \mathrm{Var}[X|A=b]$$

implies:

$$\pi\sigma_{a1}^2 + (1-\pi)\sigma_{a0}^2 > \pi\sigma_{b1}^2 + (1-\pi)\sigma_{b0}^2 \tag{12}$$

using Equation (6). Further, suppose that

$$\pi < 0.5 \quad \text{and} \quad \mu_1 \neq \mu_0.$$

Consider the following two cases that demonstrate that AUC_a can be greater than or less than AUC_b.

Case A ($\sigma_{a0}^2 = \sigma_{a1}^2 = \sigma_a^2$ and $\sigma_{b0}^2 = \sigma_{b1}^2 = \sigma_b^2$): It follows from Equation (12) that $\sigma_a^2 > \sigma_b^2$ so

$$\mathrm{AUC}_a = \Phi\left(\frac{\mu_1-\mu_0}{\sqrt{2\sigma_a^2}}\right) < \Phi\left(\frac{\mu_1-\mu_0}{\sqrt{2\sigma_b^2}}\right) = \mathrm{AUC}_b.$$

Here, we also use the fact that $\mu_1 \neq \mu_0$ and that $\Phi(\cdot)$ is a monotonically increasing function.

Case B ($\sigma_{a0}^2 = \sigma_{a1}^2 = \sigma_a^2$): It follows from Equation (12) that:

$$\sigma_a^2 > \pi\sigma_{b1}^2 + (1-\pi)\sigma_{b0}^2.$$

Let

$$\sigma_a^2 = \pi\sigma_{b1}^2 + (1-\pi)\sigma_{b0}^2 + \varepsilon \tag{13}$$

where $\varepsilon > 0$. We want to find conditions under which $\mathrm{AUC}_a \geq \mathrm{AUC}_b$. It follows from Equation (5) that $\mathrm{AUC}_a \geq \mathrm{AUC}_b$ when $2\sigma_a^2 \leq \sigma_{b1}^2 + \sigma_{b0}^2$ (since $\Phi(\cdot)$ is a monotonically increasing function). Incorporating Equation (13), the AUC condition requires:

$$\sigma_{b1}^2 + \sigma_{b0}^2 \geq 2\pi\sigma_{b1}^2 + 2(1-\pi)\sigma_{b0}^2 + 2\varepsilon,$$

which simplifies to:

$$\sigma_{b1}^2 \geq \sigma_{b0}^2 + \frac{2\varepsilon}{1-2\pi} \tag{14}$$

when $\pi < 0.5$.

Note that the smaller class needs to have higher variance for Equation (14) to hold. Class-conditional variances are weighted in the expected overall unconditional variance but not weighted in the AUC formula. The closer we are to class balance (i.e., $\pi = 0.5$) the greater the difference in class-conditional variances we need for $\mathrm{AUC}_a \geq \mathrm{AUC}_b$. □

D maxWAUC Algorithm

Procedure 2: maxWAUC (t-th iteration)

Input: data owned $(\boldsymbol{X_i}, A_i, Y_i)_{i=1}^N$, scoring algorithm r, bias threshold ε, set of acquired features $Q(t)$, data available for acquisition $(\boldsymbol{Z_i})_{i=1}^N$;
Output: $Q(t+1)$;
if $\boldsymbol{A}$ *cannot be used* **then**
  $\boldsymbol{S} \coloneqq r(\hat{\boldsymbol{X}}, \boldsymbol{Y})$;
else
  $\boldsymbol{S} \coloneqq r(\hat{\boldsymbol{X}}, \boldsymbol{A}, \boldsymbol{Y})$;
for *group* $g \in \{a, b\}$ **do**
  compute $AUC_g(S)$ (Definition 3.1) ;
Bias $\coloneqq 1 - \frac{\min_g(\mathrm{AUC}_g(S))}{\max_g(\mathrm{AUC}_g(S))}$ (Definition 3.2) ;
if Bias $> \varepsilon$ **then**
  for *feature* $\boldsymbol{Z}^j \in \hat{\boldsymbol{Z}}, j \notin Q(t)$ **do**
    if $\boldsymbol{A}$ *cannot be used* **then**
      for feature $\boldsymbol{Z}^j$ and score $\boldsymbol{S}$, obtain class-conditional means, $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$, and covariance matrices, $\boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$ (Overall Summary Statistics Subroutine);
      $h(\boldsymbol{S}, \boldsymbol{Z}^j) \coloneqq \Phi\left(\sqrt{(\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^\top (\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})}\right)$;
    else
      for $g \in \{a, b\}$ **do**
        for feature $\boldsymbol{Z}^j$, group g, and score $\boldsymbol{S}$, obtain class-conditional means, $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$, and covariance matrices, $\boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$ (Summary Statistics by Group Subroutine);
        ϕ_g represents the proportion of individuals from group g ;
        $\omega_g \coloneqq (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})^\top (\boldsymbol{\Sigma_0} + \boldsymbol{\Sigma_1})^{-1} (\boldsymbol{\mu_1} - \boldsymbol{\mu_0})$;
      $h(\boldsymbol{S}, \boldsymbol{Z}^j) \coloneqq \phi_a \Phi(\sqrt{\omega_a}) + \phi_b \Phi(\sqrt{\omega_b})$;
  $j^\star \coloneqq \arg\max_j h(\boldsymbol{S}, \boldsymbol{Z}^j)$;
  acquire feature $\boldsymbol{Z}^{j^\star}$;
  return $Q(t+1) \coloneqq Q(t) \cup \{j^\star\}$;
else
  no intervention;

Subroutine: Overall Summary Statistics

Input: feature available for acquisition $\boldsymbol{Z}$, existing score $\boldsymbol{S}$;
Output: class-conditional mean vectors $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}$, class-conditional covariance matrices $\boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$;
for *class* $y \in \{0, 1\}$ **do**
  $n \coloneqq n_{Y=y}$;
  $\boldsymbol{\mu_y} \coloneqq \begin{bmatrix} \bar{S}_y \\ \bar{Z}_y \end{bmatrix} = \begin{bmatrix} \frac{1}{n} \sum_{i:\, Y_i = y} S_i \\ \frac{1}{n} \sum_{i:\, Y_i = y} Z_i \end{bmatrix}$;
  $\boldsymbol{\Sigma_y} \coloneqq \begin{bmatrix} \sigma_{S,y}^2 & \rho_y \sigma_{S,y} \sigma_{Z,y} \\ \rho_y \sigma_{S,y} \sigma_{Z,y} & \sigma_{Z,y}^2 \end{bmatrix}$
  where $\sigma_{S,y}^2 = \frac{1}{n-1} \sum_{i:\, Y_i = y} (S_i - \bar{S}_y)^2$, $\sigma_{Z,y}^2 = \frac{1}{n-1} \sum_{i:\, Y_i = y} (Z_i - \bar{Z}_y)^2$, and $\rho_y = \frac{1}{(n-1)\sigma_{s,y}\sigma_{Z,y}} \sum_{i:\, Y_i = y} (S_i - \bar{S}_y)(Z_i - \bar{Z}_y)$;
return $\boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma_0}, \boldsymbol{\Sigma_1}$

E Plots of Synthetic and Empirical Data without Protected Attribute

E.1 Synthetic Data Analysis

We generate $N = 20,000$ individuals with 50 non-protected continuous normally distributed features, one binary protected feature (group), and a binary outcome (class). Of the 50 features, half are *informative* in that the class-conditional distributions of each of the features have means that are separated from each other. The remaining features are *uninformative* random noise features. Group a constitutes 70% and group b 30%. The base rate of positive class labels in both groups is 25%.

For the synthetic data, there is nothing fundamentally different between the two groups besides the number of individuals in each group so any difference in predictive performance stems from the feature selection procedure and the design of the algorithm (i.e., whether the attribute A is used in the classifier).

We set the level of acceptable bias $\varepsilon = 10^{-6}$ to demonstrate the various procedures over many rounds. We collect 15 additional features for classification using each one and compare the procedures when A is and is not used in the scoring function. We randomly select one feature to represent the data the manager begins with (Round 0) for classification.

Figure 5: Group AUCs over feature augmentation rounds without protected attribute

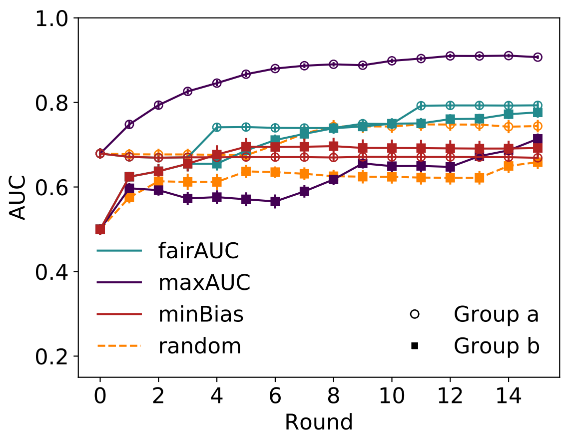


Figure 6: Accuracy-fairness tradeoff without protected attribute

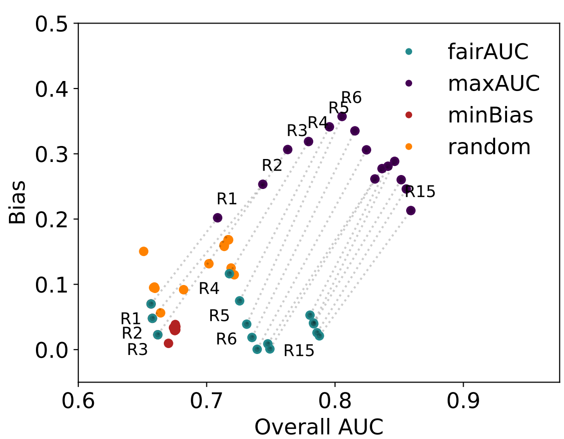
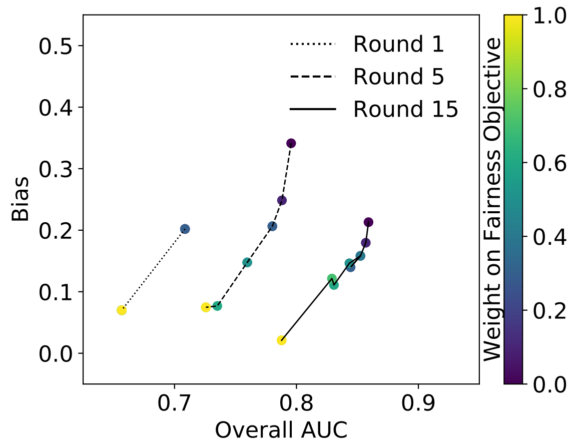


Figure 7: Pareto frontier without protected attribute

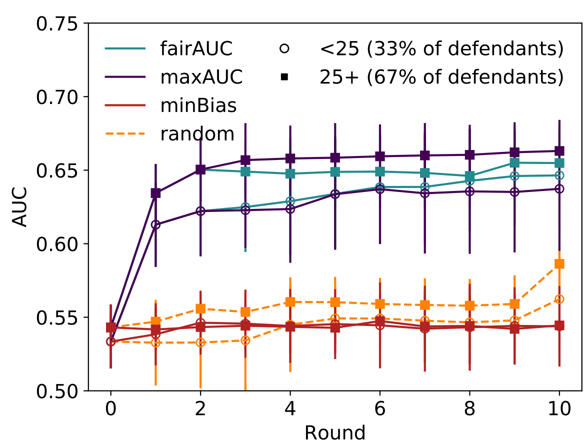


E.2 Real Data Analysis

We define age as age at charge, which is different from the age recorded in the ProPublica dataset. ProPublica records defendants' age in 2016, the year the data was collected, rather than the age at charge. We calculate age at charge by subtracting date of birth from the date the defendant went to jail.

.

Figure 8: Predicting violent recidivism without protected attribute (age)



F Effect of fairAUC on the AUC of Each Group

In this section, we analyze the fairAUC procedure in the binormal framework for features Su and Liu (1993) (where the features follow a normal distribution conditioned on the class and the protected group). We show that if fairAUC uses FLD-based scores S (Equation (15)), then in each iteration $t \in \mathbb{N}$, where the AUC for disadvantaged group is bounded away from 1 and there is at least one auxiliary feature which has "low" class-conditional covariances with the current scores S and has "bounded" class-conditional variances and means, fairAUC improves the AUC of the disadvantaged group by at least a constant in iteration t (Theorem F.4).

From Section 4.4, recall that there are a total of m features, out of which, the decision-maker initially has access to d *acquired features*:

$$X := (X^1, X^2, \ldots, X^d) \in \mathbb{R}^d, \qquad \text{(Acquired features)}$$

and has the option to augment $d' \coloneqq m - d$ *auxiliary features*:

$$Z \coloneqq (Z^1, Z^2, \ldots, Z^{d'}) \in \mathbb{R}^{d'}. \qquad \text{(Auxiliary features)}$$

The m features $X \cup Z$, together with class label Y and group A, are assumed to follow the following binormal framework in this section.

Definition F.1 (**Binormal framework**). *The m features $X \cup Z$, the class label Y, and the group label A are distributed according to a distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathbb{R}^{d'} \times \{0,1\} \times \{a,b\}$, such that for each $y \in \{0,1\}$ and $g \in \{a,b\}$, conditioned on $A = g$ and $Y = y$, the m features, follow a m-variate normal distribution with an invertible covariance matrix. (Note that, conditioned on $A = g$ and $Y = y$ different features can be correlated with each other.)*

From the distribution $\mathcal{D}$ (in Definition F.1), $N \in \mathbb{N}$ independent samples are drawn to construct a dataset D before starting the fairAUC procedure. The summary statistics subroutines (Subroutines SSR and SSR2) use D, every time they are queried, to compute the approximations to first and second moments of the distribution $\mathcal{D}$; we assume that these approximations have a negligible error (Assumption 1).

Assumption 1. *Assume that the sample means and covariances computed by two summary statistics subroutines (Subroutines SSR and SSR2) are equal to the corresponding true means and covariances of draws from $\mathcal{D}$.*

Since the subroutines use independent samples from $\mathcal{D}$, where the features follow a normal distribution, from the concentration inequalities of the normal distribution Tropp (2015), we expect the samples means and covariances of the features on D to be "good approximations" of the true means and covariances of the features on $\mathcal{D}$ for large N.

At each iteration, fairAUC acquires one auxiliary feature. For each $t \in [d']$, let $Q(t) \subseteq [d']$ denote the set of all auxiliary features acquired before the start of the t-th iteration; where we have $Q(1) \coloneqq \emptyset$. Further, let $X(t)$ denote the tuple of all features in $Q(t)$ and the d features $(X^1, X^2, \ldots, X^d)$, i.e.,

$$X(t) \coloneqq (X^1, X^2, \ldots, X^d) \ \cup \ \left(Z^\ell\right)_{\ell \in Q(t)}.$$

Note that as $Q(1) = \emptyset$, $X(1) = X$.

We need to define the AUC of $X(t)$ for a group $g \in \{a,b\}$ (Definition F.3) before stating our results. Note that $X(t)$ is always of the form $X \cup \{Z\}_{\ell \in Q(t)}$, i.e., $k \geq 0$ auxiliary features augmented to X. We restrict our definition of the AUC (Definition F.3) to such sets of features.

Definition F.2 (**AUC of linear classifiers on group g**). *Given $k \geq 0$ auxiliary features, say $Z^1, Z^2, \ldots, Z^k$, acquired features $X \in \mathbb{R}^d$, a vector $w \in \mathbb{R}^{d+k}$, and a group $g \in \{a,b\}$, consider a classifier C that given threshold $\tau \in \mathbb{R}$, predicts $\mathbb{I}[\sum_{i=1}^d w_i X^i + \sum_{i=1}^k w_{d+i} Z^i > \tau]$. Then the AUC of C for group g, denoted by*

$$\mathrm{AUC}_g(w, X, Z^1, \ldots, Z^k),$$

is the area under the ROC curve of C when samples $((X,Z),Y,A)$ are drawn from $\mathcal{D}$ conditioned on $A = g$.

Definition F.3 (**AUC for group g**). *Given $k \geq 0$ auxiliary features, say $Z^1, Z^2, \ldots, Z^k$, acquired features $X \in \mathbb{R}^d$, and a group $g \in \{a,b\}$, define the AUC of $(X, Z^1, \ldots, Z^k)$ for group g as*

$$\mathrm{AUC}_g(X, Z^1, \ldots, Z^k) \coloneqq \max_{w \in \mathbb{R}^{d+k}} \mathrm{AUC}_g(w, X, Z^1, \ldots, Z^k).$$

Using Definition F.3, we can formalize the "FLD-based" score that fairAUC uses in this section. For all samples in group $g \in \{a,b\}$ (i.e., $i \in [N]$, with $A = g$), define the scores $S(t) \in \mathbb{R}$ used in fairAUC as the projection of $X(t)$ that maximizes the AUC of the resulting linear classifier on group g (see Definition F.2):

$$S(t) \coloneqq \langle w^\star, X(t) \rangle, \ \text{where } w^\star \coloneqq \mathrm{argmax}_{w \in \mathbb{R}^{d+t}} \mathrm{AUC}_{g(t)}(w, X(t)). \qquad (15)$$

One can show that $S(t)$ is equivalent to the projection obtained using FLD on each group; this uses the fact that the data follows the binormal framework (Definition F.1; see Su and Liu (1993)).

Procedure: fairAUC (t-th iteration)

Input: Data owned $(X_i, A_i, Y_i)_{i=1}^N$, $[d']$, indices acquired $Q(t) \subseteq [d']$, and data acquired $(Z_i^\ell)_{i\in[N],\ \ell\in Q(t)}$;
Output: Set $Q(t+1) \subseteq [d']$ of the auxiliary features augmented;

for *group* $g \in \{a, b\}$ **do**
    // Compute FLD scores
    Query $\Sigma_0^{(g)}, \Sigma_1^{(g)}, \mu_0^{(g)}, \mu_1^{(g)} = \mathtt{SSR}(X \cup (Z^\ell)_{\ell\in Q(t)}, A, g)$;
    Compute $\Delta\mu^{(g)} \coloneqq (|\mu_{11}^{(g)} - \mu_{01}^{(g)}|, \ldots, |\mu_{1d}^{(g)} - \mu_{0d}^{(g)}|)$;
    Compute $\Sigma^{(g)} \coloneqq \Sigma_0^{(g)} + \Sigma_1^{(g)}$;
    Initialize $S \coloneqq (0)_{i=1}^N$;
    for $i \in [N]$ **do**
        if $A_i = a$ **then**
            Set $S_i \coloneqq (\Delta\mu^{(a)})^\top (\Sigma^{(a)})^{-1} X_i$;
        else
            Set $S_i \coloneqq (\Delta\mu^{(b)})^\top (\Sigma^{(b)})^{-1} X_i$;

// Identify disadvantaged group
for *group* $g \in \{a, b\}$ **do**
    Compute $\mathrm{AUC}_g(X) \coloneqq \Phi\left(\sqrt{(\mu_1^{(g)} - \mu_0^{(g)})^\top (\Sigma_0^{(g)} + \Sigma_1^{(g)})^{-1} (\mu_1^{(g)} - \mu_0^{(g)})}\right)$;

$g(t) \coloneqq \arg\min_{g\in\{a,b\}}(\mathrm{AUC}_g(X))$ // Find disadvantaged group ;

for *auxiliary feature* $\ell \in [d']$ **do**
    // For group $g(t)$ query: class-conditional means $\mu_0, \mu_1 \in \mathbb{R}^2$, and
    // covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{R}^{2\times 2}$ between score S and auxiliary feature Z^ℓ.
    Query $\Sigma_0, \Sigma_1, \mu_0, \mu_1 = \mathtt{SSR2}(\ell, g(t), S)$;
    Compute $\mathrm{AUC}_{g(t)}(S, Z^\ell) \coloneqq \Phi\left(\sqrt{(\mu_1 - \mu_0)^\top (\Sigma_0 + \Sigma_1)^{-1} (\mu_1 - \mu_0)}\right)$;

$i \coloneqq \arg\max_{\ell\in[d']} \mathrm{AUC}_{g(t)}(S, Z^\ell)$;
$Q(t+1) = Q(t) \cup \{i\}$;

return $Q(t+1)$.

Now, we restrict out attention to a particular iteration $t \in \mathbb{N}$. We define certain quantities that show up in our results (Theorem F.4). Suppose $g(t) \in \{a, b\}$ is the disadvantaged group in the t-th iteration. For each auxiliary feature $\ell \in [d']\backslash Q(t)$, let $\Delta v_\ell^{(t)}$ be the absolute difference of its class conditional means (on $g(t)$), i.e.,

$$\Delta v_\ell^{(t)} \coloneqq \left|\mathbb{E}[Z^\ell \mid Y = 1, A = g(t)] - \mathbb{E}[Z^\ell \mid Y = 0, A = g(t)]\right|. \tag{16}$$

Subroutine: `SSR` (summary statistic subroutine)

Input: Acquired features $\{X_i^j\}_{i\in[N], j\in[d+t]}$, protected attributes $\{A_i\}_{i=1}^N$, group $g \in \{a, b\}$;
Output: class-conditional mean vectors $\mu_0, \mu_1 \in \mathbb{R}^d$, class-conditional covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{R}^{d\times d}$;
for *class* $y \in \{0, 1\}$ **do**
    Compute $n \coloneqq \sum_i \mathbb{I}[A_i = g, Y_i = y]$ // Total elements with $A_i = g$ and $Y_i = y$
    Compute $\mu_y \coloneqq \frac{1}{n}\left[\sum_{i\colon A_i=g, Y_i=y} X_i^1, \ldots, \sum_{i\colon A_i=g, Y_i=y} X_i^d\right]$ // Empirical-mean of X when $A_i = g$ and $Y_i = y$
    Compute matrix $\Sigma_y \in \mathbb{R}^{d\times d}$, where for all $\ell, k \in [d]$ // Empirical-covariance matrix of X when $A_i = g$ and $Y_i = y$

$$(\Sigma_y)_{\ell,k} \coloneqq \frac{1}{n-1} \sum_{i\colon A_i=g, Y_i=y} (X_i^\ell - (\mu_y)_\ell)(X_i^k - (\mu_y)_k).$$

return $\mu_0, \mu_1, \Sigma_0, \Sigma_1$

Subroutine: SSR2 (Summary statistic subroutine - 2)

Input: auxiliary feature index $\ell \in [d']$, group g, score $\{S_i\}_{i=1}^N$ (Also, has access to all auxiliary features $\{Z_i\}_{i=1}^N$);

Output: Class-conditional mean vectors $\mu_0, \mu_1 \in \mathbb{R}^2$, class-conditional covariance matrices $\Sigma_0, \Sigma_1 \in \mathbb{R}^{2\times 2}$;

for *class* $y \in \{0,1\}$ **do**

  Compute $n \coloneqq \sum_i \mathbb{I}[A_i = g, Y_i = y]$ // Total elements with $A_i = g$ and $Y_i = y$

  Compute $\mu_{S,y} \coloneqq \frac{1}{n}\sum_{i\colon A_i=g, Y_i=y} S_i$

  Compute $\mu_{Z,y} \coloneqq \frac{1}{n}\sum_{i\colon A_i=g, Y_i=y} Z_i^\ell$

  Compute $\Sigma_y \coloneqq \begin{bmatrix} \sigma_{S,y}^2 & \rho_y \\ \rho_y & \sigma_{Z,y}^2 \end{bmatrix}$ where

$$\sigma_{S,y}^2 \coloneqq \frac{1}{n-1}\sum_{i\colon A_i=g, Y_i=y} (S_i - \mu_{S,y})^2,$$
$$\sigma_{Z,y}^2 \coloneqq \frac{1}{n-1}\sum_{i\colon A_i=g, Y_i=y} (Z_i^\ell - \mu_{Z,y})^2, \text{ and}$$
$$\rho_y \coloneqq \frac{1}{n-1}\sum_{i\colon A_i=g, Y_i=y} (S_i - \mu_{S,y}) \cdot (Z_i^\ell - \mu_{Z,y})$$

return $\mu_0, \mu_1, \Sigma_0, \Sigma_1$

Next, using $\Delta v_\ell^{(t)}$ define the following quantities for each auxiliary feature $\ell \in [d']\backslash Q(t)$

$$\beta_\ell^{(t)} \coloneqq \frac{\Delta v_\ell^{(t)}}{\sqrt{\sum_{y\in\{0,1\}} \mathrm{Var}[Z^\ell \mid Y = y, A = g(t)]}}, \tag{17}$$

$$\delta_\ell^{(t)} \coloneqq \frac{1}{\Delta v_\ell^{(t)}} \cdot \left|\sum_{y\in\{0,1\}} \mathrm{Cov}[S(t), Z^\ell \mid Y = y, A = g(t)]\right|, \tag{18}$$

Finally, let define $\gamma^{(t)}$ as

$$\gamma^{(t)} \coloneqq 1 - \mathrm{AUC}_{g(t)}(X(t)).$$

We prove Theorem F.4.

Theorem F.4 (Effect of fairAUC on the AUC of the disadvantaged group). *Suppose that the m features $X \cup Z$, class label Y, and protected group A follow the binormal framework (Definition F.1). Further, assume that two summary statistics subroutines satisfy Assumption 1. Then, for all iterations $t \in [d']$ and all auxiliary features $\ell \in [d']\backslash Q(t)$, it holds that*

$$\mathrm{AUC}_{g(t)}(S(t), Z^\ell) - \mathrm{AUC}_{g(t)}(X(t)) \quad > \quad \frac{1}{18} \cdot \left(\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)})\right)^2. \quad \text{(AUC increment on selecting } Z^\ell\text{)}$$

Further, the auxiliary feature $i \in [d']\backslash Q(t)$ selected by fairAUC in the t-th iteration satisfies

$$\mathrm{AUC}_{g(t)}(X(t), Z^i) - \mathrm{AUC}_{g(t)}(X(t)) \quad \geq \quad \max_{\ell\in[d']\backslash Q(t)} \frac{1}{18} \cdot \left(\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)})\right)^2. \tag{AUC increment by fairAUC; 19}$$

Some remarks are in order:

1. **(Dependence on $\gamma^{(t)}$).** As $\mathrm{AUC}_{g(t)}(X(t))$ approaches 1 (i.e., $\gamma^{(t)}$ approaches 0), the lower bound in Equation (19) approaches 0. This is expected because when $\mathrm{AUC}_{g(t)}(X(t))$ is close to 1, which is its maximum value, each auxiliary feature can only increment the AUC for $g(t)$ by a small amount.

2. **(Dependence on $\beta_\ell^{(t)}$).** If $\Delta|v_\ell^{(t)}|$ is small or $\sum_{y\in\{0,1\}} \mathrm{Var}[Z^\ell \mid Y = y, A = g(t)]$ is large, then Equation (5) tells us the classifier which uses Z^ℓ to predict the class Y has a low AUC, i.e., Z^ℓ is not a "good predictor" of Y. This is captured by $\beta_\ell^{(t)}$ in Equation (19). To see this, observe that when $\Delta|v_\ell^{(t)}|$ is small or $\sum_{y\in\{0,1\}} \mathrm{Var}[Z^\ell \mid Y = y, A = g(t)]$ is large, $\beta_\ell^{(t)}$ is small. Thus, the increment in the AUC is also small.

3. **(Dependence on $\delta_\ell^{(t)}$).** To gain some intuition about the dependence on $\delta_\ell^{(t)}$, consider the extreme case, where Z^ℓ is identical to $S(t)$. This maximizes the class-conditional covariances of Z^ℓ and $S(t)$ (on $A = g(t)$) subject to a fixed value of variance of Z^ℓ. Thus, it also maximizes $\delta_\ell^{(t)}$. However, in this case, any linear combination of $X(t)$ and Z^ℓ is identical to some linear combination of $X(t)$ (and vice-versa).[6] Thus, $\mathrm{AUC}_{g(t)}(X(t), Z^\ell) = \mathrm{AUC}_{g(t)}(X(t))$. Intuitively, Z^ℓ does not provide any new information.

Theorem F.4 shows the effect of fairAUC on the AUC of the disadvantaged group. Our next result (Theorem F.5), captures the effect of fairAUC on the AUC of the advantaged group. Suppose $\hat{g}(t) \in \{a, b\}$ be the advantaged group at the t-th iteration. Theorem F.5 provides a lower bound in the improvement on the AUC of the advantaged group in the t-th iteration in terms of quantities $\Delta\hat{v}_\ell^{(t)}$, $\hat{\beta}_\ell^{(t)}$, $\hat{\delta}_\ell^{(t)}$, and $\hat{\gamma}^{(t)}$ (Equations (20) to (23)); these are equivalent to $\Delta v_\ell^{(t)}$, $\beta_\ell^{(t)}$, $\delta_\ell^{(t)}$, and $\gamma^{(t)}$ in Theorem F.4, except the disadvantaged group $g(t)$ in the definitions changes to the advantaged group $\hat{g}(t)$.

Formally, we define $\Delta\hat{v}_\ell^{(t)}$, $\hat{\beta}_\ell^{(t)}$, $\hat{\delta}_\ell^{(t)}$, and $\hat{\gamma}^{(t)}$ as follows.

$$\Delta\hat{v}_\ell^{(t)} \coloneqq \left|\mathbb{E}[Z^\ell \mid Y = 1, A = \hat{g}(t)] - \mathbb{E}[Z^\ell \mid Y = 0, A = \hat{g}(t)]\right|, \tag{20}$$

$$\hat{\beta}_\ell^{(t)} \coloneqq \frac{\Delta v_\ell^{(t)}}{\sqrt{\sum_{y\in\{0,1\}} \mathrm{Var}[Z^\ell \mid Y = y, A = \hat{g}(t)]}}, \tag{21}$$

$$\hat{\delta}_\ell^{(t)} \coloneqq \frac{1}{\Delta v_\ell^{(t)}} \cdot \left|\sum_{y\in\{0,1\}} \mathrm{Cov}[S(t), Z^\ell \mid Y = y, A = \hat{g}(t)]\right|, \tag{22}$$

$$\hat{\gamma}^{(t)} \coloneqq 1 - \mathrm{AUC}_{\hat{g}(t)}(X(t)). \tag{23}$$

Theorem F.5 (**Effect of fairAUC on the AUC of the advantaged group**). *Suppose that the m features $X \cup Z$, class label Y, and protected group A follow the binormal framework (Definition F.1). Further, assume that two summary statistics subroutines satisfy Assumption 1. Then, for all iterations $t \in [d']$ the auxiliary feature $i \in [d']\backslash Q(t)$ selected by fairAUC in the t-th iteration satisfies*

$$\mathrm{AUC}_{\hat{g}(t)}(X(t), Z^i) - \mathrm{AUC}_{\hat{g}(t)}(X(t)) \quad \geq \quad \frac{1}{18} \cdot \left(\hat{\gamma}^{(t)} \cdot \hat{\beta}_i^{(t)} \cdot (1 - \hat{\delta}_i^{(t)})\right)^2. \quad \text{(AUC increment by fairAUC; 24)}$$

At a first glance, the lower bound in Theorem F.5 may appear to be equivalent to Theorem F.4. The difference is that, the fairAUC is guaranteed to pick the "best" feature for the disadvantaged group, but it may not pick the best feature for the minority group. Thus, while the lower bound in Theorem F.4 is at large if any auxiliary feature $\ell \in [d']\backslash Q(t)$, has large $\beta_\ell^{(t)}$ and small $\delta_\ell^{(t)}$, whereas Theorem F.5 requires $\beta_i^{(t)}$ to be large and $\delta_i^{(t)}$ to be small for the particular feature $i \in [d']\backslash Q(t)$, selected by fairAUC.

F.1 Preliminaries

In this section, we present three lemmas which will be used in proof of Theorem F.4.

Lemma F.6 (**Expression for optimal AUC (Su and Liu 1993, Corollary 3.1)**). *Consider two random variables $X \in \mathbb{R}^d$ and $Y \in \{0, 1\}$, which are distributed according to a joint distribution $\mathcal{D}$, such that for all $y \in \{0, 1\}$, conditioned on $Y = y$, X follows a multivariate normal distribution with mean $\mu_y \in \mathbb{R}^d$ and covariance matrix $\Sigma_y \in \mathbb{R}^{d\times d}$:*

$$\text{for all } y \in \{0, 1\}, \quad X \mid Y = y \quad \sim \quad \mathcal{N}(\mu_y, \Sigma_y).$$

Let $\Delta\mu \coloneqq |\mu_1 - \mu_0|$ and $\Sigma \coloneqq \Sigma_0 + \Sigma_1$. Then, the maximum AUC of a linear classifier which takes X as input and predicts Y is $\Phi\left(\sqrt{\Delta\mu\Sigma^{-1}\Delta\mu}\right)$.

Lemma F.7 (**"Inverting" $\Phi(\sqrt{\cdot})$**). *For all $\alpha \geq 0$ and $\gamma > 0$, if $\Phi(\sqrt{\alpha}) < 1 - \gamma$ then it holds that $\alpha < 2 \cdot \ln(1/\gamma)$.*

[6] This uses the fact that $S(t)$ is a linear combination of $X(t)$. Since Z^ℓ is identical to $S(t)$, Z^ℓ is also linear combination of $X(t)$.

Proof. Proof. We use the fact that for all $x \in \mathbb{R}$, the inequality $\Phi(x) \geq 1 - e^{-x^2/2}$ holds (see, e.g., (Roch 2014, Equation 2.24)). Applying this, we get

$$\Phi\left(\sqrt{\alpha}\right) \geq 1 - e^{-\alpha/2}.$$

Chaining the above inequality with $1 - \gamma > \Phi\left(\sqrt{\alpha}\right)$ and rearranging, we get $\alpha < 2 \cdot \ln(1/\gamma)$. $\square$

Lemma F.8 (Lower bound on change in $\Phi(\sqrt{\cdot})$). *For all $\gamma > 0$, $\Delta_0 > 0$, $\alpha \in (0, 2 \cdot \ln(1/\gamma))$ and $\Delta \geq \Delta_0$, it holds that*

$$\Phi(\sqrt{\alpha + \Delta}) - \Phi(\sqrt{\alpha}) \geq \frac{\gamma^2 \cdot \Delta_0}{6 \cdot (1 + \Delta_0)^{3/2}}. \tag{25}$$

We note that the bound in Lemma F.8 weakens as Δ_0 (and so, Δ) increases. To see this, observe that the LHS in Equation (25) is an increasing function of Δ. In contrast, if Δ_0 is large enough, the RHS in Equation (25) is a decreasing function of Δ_0. Nevertheless, Lemma F.8 suffices to prove Theorem F.4.

The proof of Lemma F.8 appears in Section F.4.

F.2 Proof of Theorem F.4

In this section, we present a proof of Theorem F.4. We begin with the necessary notation and the lemmas (Lemmas F.9 and F.10) in Section F.2.1. Next, in Section F.2.2, we complete the proof of Theorem F.4 assuming Lemmas F.9 and F.10. Finally, in Sections F.2.3 and F.2.4, we present the proofs of Lemmas F.9 and F.10 respectively.

Recall that we are given distribution $\mathcal{D}$ which satisfies Definition F.1. We assume that the statistics returned by the two summary statistic subroutines are exact (Assumption 1).

Fix any iteration $t \in [d']$. Our goals are to prove that for each auxiliary feature $\ell \in [d'] \backslash Q(t)$, $\mathrm{AUC}_{g(t)}(S(t), Z^\ell) - \mathrm{AUC}_{g(t)}(X) \geq \frac{1}{18} \cdot (\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)}))^2$, and that, in this iteration, fairAUC improves the AUC for the current disadvantaged group $g(t)$, by at least

$$\mathrm{AUC}_{g(t)}(X, Z^i) - \mathrm{AUC}_{g(t)}(X) \quad > \quad \max\nolimits_{\ell \in [d']} \frac{1}{18} \cdot \left(\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)})\right)^2.$$

Fix any auxiliary feature $\ell \in [d']$. Then, the proof proceeds in two broad steps. First, we show that $\mathrm{AUC}(X, Z^\ell)$ is lower bounded by $\mathrm{AUC}(S(t), Z^\ell)$ (Lemma F.9). Then, we derive an explicit formula and lower bound for $\mathrm{AUC}(S(t), Z^\ell)$ (Lemma F.10). This formula is the same as the formula used to compute $\mathrm{AUC}_{g(t)}(S(t), Z^\ell)$ in fairAUC. Thus, fairAUC selects the auxiliary feature i where

$$i \in \operatorname{argmax}_{\ell \in [d']} \mathrm{AUC}_{g(t)}(S(t), Z^\ell).$$

Combining this with a lower bound $\mathrm{AUC}(S(t), Z^\ell)$ for any $\ell \in [d']$, we get that the auxiliary feature i selected by fairAUC, satisfies

$$\mathrm{AUC}_{g(t)}(S(t), Z^i) > \max\nolimits_{\ell \in [d']} \frac{1}{18} \cdot \left(\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)})\right)^2.$$

Then, using Lemma F.9, we get that the auxiliary feature selected by fairAUC improves the AUC for $g(t)$ by at least $\max_{\ell \in [d']} \frac{1}{18} \cdot \left(\gamma^{(t)} \cdot \beta_\ell^{(t)} \cdot (1 - \delta_\ell^{(t)})\right)^2$. Finally since the choice of t was arbitrary, we get that the result holds for all $t \in [d']$.

F.2.1 Additional Notation and Supporting Lemmas

We begin by presenting the two lemmas used in proof of Theorem F.4.

Lemma F.9 (Projection does not increase AUC). *Consider three random variables $X \in \mathbb{R}^d$, $Y \in \{0, 1\}$, $Z \in \mathbb{R}$, which follow some joint distribution $\mathcal{D}$. Given $w \in \mathbb{R}^d$, define $S \coloneqq \langle w, X \rangle$, then it holds that*

$$\mathrm{AUC}_{\mathcal{D}}(X, Z) \geq \mathrm{AUC}_{\mathcal{D}}(S, Z).$$

The proof of Lemma F.9 appears in Section F.2.4.

We require some additional notation to presnt Lemma F.10. Fix any iteration $t \in [d']$. Since t will be fixed for the remainder of the proof, we drop the superscripts from $\gamma^{(t)}$, $\Delta v_\ell^{(t)}$, $\beta_\ell^{(t)}$, and $\delta_\ell^{(t)}$. From Definition F.1, we know that conditioned on A and Y, $X \cup Z$ follow a m-variate normal distribution. It follows that $X(t)$ also has a Gaussian distribution conditioned on Y and A (see e.g., (Stirzaker 2003, Theorem 5, Section 8.4)). Suppose for all $y \in \{0,1\}$

$$X(t) \mid Y = y, A = g(t) \quad \sim \quad \mathcal{N}(\mu_y, \Sigma_y), \qquad \text{(Binormality of acquired features; 26)}$$

and for all $y \in \{0,1\}$ and $\ell \in [d']\backslash Q(t)$,

$$Z^\ell \mid Y = y, A = g(t) \quad \sim \quad \mathcal{N}(v_{y\ell}, \sigma_{y\ell}^2), \qquad \text{(Binormality of auxiliary features; 27)}$$

where $\mu_0, \mu_1 \in \mathbb{R}^{d+t}$, $\Sigma_0, \Sigma_1 \in \mathbb{R}^{(d+t)\times(d+t)}$, and for all $\ell \in [d']\backslash Q(t)$, $v_{0\ell}, v_{1\ell} \in \mathbb{R}$ and $\sigma_{0\ell}^2, \sigma_{1\ell}^2 \geq 0$. Note that $\{Z^\ell\}_{\ell \in [d']\backslash Q(t)}$ can be correlated with each other (and with $X(t)$).

Next, we show that $\Sigma_0 + \Sigma_1$ is invertible. Towards this, notice that requires Definition F.1 that for any $y \in \{0,1\}$ and $g \in \{a,b\}$, conditioned on $Y = y$ and $A = g$ covariance matrix of all m features, say M_{yg}, is invertible. Since covariance matrices are positive semi-definite (PSD) and any invertible PSD matrix is positive definite (PD), it follows that M_{yg} is PD. Notice that Σ_0 and Σ_1 are submatrices of M_{0g} and M_{1g}. Since submatrices of PD matrices are also PD, it follows that Σ_0 and Σ_1 are PD. Then, $\Sigma_0 + \Sigma_1$ is PD. Thus, $\Sigma_0 + \Sigma_1$ is invertible.

Define $\Delta\mu$ and $\Delta\mu_S$ to be the difference in class-conditional means of $X(t)$ and S respectively

$$\begin{aligned} \Delta\mu &:= \left|\mu_1 - \mu_0\right|. \qquad (28)\\ \Delta\mu_S &:= \left|\mathbb{E}[S(t) \mid Y = 1, A = g(t)] - \mathbb{E}[S(t) \mid Y = 0, A = g(t)]\right|. \end{aligned}$$

Similarly, for all $\ell \in [d']$, define Δv_ℓ to be the difference in class-conditional means of Z^ℓ

$$\Delta v_\ell := \left|v_{1\ell} - v_{0\ell}\right|,. \qquad (29)$$

For all $y \in \{0,1\}$, define σ_{yS}^2 to be the variance of $S(t)$ conditioned on $Y = y$:

$$\sigma_{yS}^2 := \mathrm{Var}[S(t) \mid Y = y, A = g(t)].$$

Finally, for all $y \in \{0,1\}$ and $\ell \in [d']$, define $\rho_{y\ell}$ as the covariance of $S(t)$ and Z^ℓ,

$$\rho_{y\ell} := \mathrm{Cov}[S(t), Z^\ell \mid Y = y, A = g(t)].$$

Using the definition of $S(t)$ (Equation (15)), we can compute $\Delta\mu_S$ in terms of $\Delta\mu$.

$$\begin{aligned} \Delta\mu_S &:= \left|\mathbb{E}[S(t) \mid Y = 1, A = g(t)] - \mathbb{E}[S(t) \mid Y = 0, A = g(t)]\right| \\ &\overset{(15)}{=} \left|\mathbb{E}\left[\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}X(t) \mid Y = 1, A = g(t)\right] - \mathbb{E}\left[\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}X(t) \mid Y = 0, A = g(t)\right]\right| \\ &= \left|\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\left(\mathbb{E}\left[X(t) \mid Y = 1, A = g(t)\right] - \mathbb{E}\left[X(t) \mid Y = 1, A = g(t)\right]\right)\right| \\ &\qquad \text{(Linearity of expectation)} \\ &\overset{(28)}{=} \left|\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu\right| \\ &= \Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu \qquad \text{(Using that } (\Sigma_0+\Sigma_1)^{-1} \text{ is a PD matrix.; 30)} \end{aligned}$$

Similarly, for all $y \in \{0,1\}$, we can also compute σ_{yS}^2. For $y = 1$, we have

$$\begin{aligned} \sigma_{1S}^2 &:= \mathrm{Var}[S(t) \mid Y = 1, A = g(t)] \\ &\overset{(15)}{=} \mathrm{Var}\left[\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}X(t) \mid Y = 1, A = g(t)\right] \\ &= \Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Sigma_1(\Sigma_0+\Sigma_1)\Delta\mu. \qquad (31) \end{aligned}$$

In the last equality, we use the fact that for any vector $w \in \mathbb{R}^{d+t}$ and random variable $X(t) \in \mathbb{R}^{d+t}$ with covariance matrix $\Sigma \in \mathbb{R}^{(d+t)\times(d+t)}$, it holds that $\mathrm{Var}[\langle w, X(t)\rangle] = w^\top \Sigma w$. Similarly for $y = 0$ we have that

$$\sigma_{0S}^2 \quad = \quad \Delta\mu^\top(\Sigma_0 + \Sigma_1)^{-1}\Sigma_0(\Sigma_0 + \Sigma_1)\Delta\mu. \tag{32}$$

Combining Equations (31) and (32), we get

$$\sigma_{0S}^2 + \sigma_{1S}^2 \overset{(31),(32)}{=} \Delta\mu^\top(\Sigma_0 + \Sigma_1)^{-1}\Delta\mu. \tag{33}$$

Lemma F.10. *If $\mathcal{D}$ satisfies Equations (26) and (27), then it holds that*

$$\mathrm{AUC}_{g(t)}(S(t), Z) := \Phi\left(\sqrt{\begin{bmatrix}\Delta\mu_S & \Delta v_\ell\end{bmatrix}\begin{bmatrix}\sigma_{0S}^2 + \sigma_{1S}^2 & \rho_{0\ell} + \rho_{1\ell} \\ \rho_{0\ell} + \rho_{1\ell} & \sigma_{0\ell}^2 + \sigma_{1\ell}^2\end{bmatrix}^{-1}\begin{bmatrix}\Delta\mu_S \\ \Delta v_\ell\end{bmatrix}}\right). \tag{34}$$

Further, let $\alpha \geq 0$, be such that $\mathrm{AUC}_{g(t)}(X) = \Phi\left(\sqrt{\alpha}\right)$, then Equation (34) implies that

$$\mathrm{AUC}_{g(t)}(S(t), Z) \geq \Phi\left(\sqrt{\alpha + \frac{\left(\Delta v_\ell - (\rho_{0\ell} + \rho_{1\ell})\right)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2}}\right). \tag{35}$$

The proof of Lemma F.10 appears in Section F.2.3. Define $\alpha \geq 0$ to be a constant, such that

$$\mathrm{AUC}_{g(t)}(X(t)) = \Phi\left(\sqrt{\alpha}\right). \tag{36}$$

(α is uniquely defined as $\Phi(\sqrt{\cdot})$ is a strictly increasing function.) Further, define $\Delta' \in \mathbb{R}$ to be term added to α in Equation (35):

$$\Delta' := \frac{\left(\Delta v - (\rho_{0\ell} + \rho_{1\ell})\right)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2} \tag{37}$$

F.2.2 Proof of Theorem F.4

Now, we are ready to complete the proof of Theorem F.4. Fix any auxiliary feature $\ell \in [d']$. Consider two cases depending on whether $\rho_{0\ell} + \rho_{1\ell} < \Delta v_\ell$.

Case A ($|\rho_{0\ell} + \rho_{1\ell}| < \Delta v_\ell$): In this case, from Equation (18), we have that $\delta_\ell = \Delta v_\ell^{-1} \cdot |\rho_{0\ell} + \rho_{1\ell}| \in (0, 1)$. Thus,

$$\Delta' = \frac{\left(\Delta v_\ell - (\rho_{0\ell} + \rho_{1\ell})\right)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2} \geq \frac{\Delta v_\ell^2 \cdot (1 - \delta_\ell)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2}.$$

Case B ($|\rho_{0\ell} + \rho_{1\ell}| \geq \Delta v_\ell$): In this case, from Equation (18), we have that $\delta_\ell = 1$. Thus,

$$\Delta' = \frac{\left(\Delta v_\ell - (\rho_{0\ell} + \rho_{1\ell})\right)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2} \geq 0 = \frac{\Delta v_\ell^2 \cdot (1 - \delta_\ell)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2}.$$

Combining both cases and using Equation (17), we can lower bound Δ' (defined in Equation (37)) as follows

$$\Delta' = \frac{\left(\Delta v_\ell - (\rho_{0\ell} + \rho_{1\ell})\right)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2} \overset{\text{(Cases A and B)}}{\geq} \frac{\Delta v_\ell^2 \cdot (1 - \delta_\ell)^2}{\sigma_{0\ell}^2 + \sigma_{1\ell}^2} \overset{(17)}{\geq} {\beta_\ell}^2 \cdot (1 - \delta_\ell)^2. \tag{38}$$

Define Δ_0 as the RHS of the above equation, i.e., $\Delta_0 := {\beta_\ell}^2 \cdot (1 - \delta_\ell)^2$. Then, we can rewrite Inequality (38) as

$$\Delta' \geq \Delta_0. \tag{39}$$

Using Lemma F.10, we can show a lower bound on the improvement in the AUC

$$\begin{aligned}
\mathrm{AUC}_{g(t)}(S(t), Z^\ell) - \mathrm{AUC}_{g(t)}(X(t)) &\overset{(36)}{=} \mathrm{AUC}_{g(t)}(S(t), Z) - \Phi\left(\sqrt{\alpha}\right) \\
&\overset{(37),\text{Lemma F.10}}{\geq} \Phi\left(\sqrt{\alpha + \Delta'}\right) - \Phi\left(\sqrt{\alpha}\right) \\
&\geq \frac{\gamma^2 \cdot \Delta_0}{6(1+\Delta_0)^{3/2}}. \\
&\qquad \text{(Using Equation (39), Lemma F.7, and Lemma F.8; 40)} \\
&= \frac{\gamma^2 \cdot {\beta_\ell}^2 \cdot (1-\delta_\ell)^2}{6(1+{\beta_\ell}^2 \cdot (1-\delta_\ell)^2)^{3/2}} \\
&\qquad \text{(Substituting } \Delta_0 \coloneqq {\beta_\ell}^2 \cdot (1-\delta_\ell)^2\text{)} \\
&\geq \frac{\gamma^2 \cdot {\beta_\ell}^2 \cdot (1-\delta_\ell)^2}{6 \cdot 2^{3/2}} \qquad \text{(Using } 0 \leq \delta_\ell, \beta_\ell \leq 1\text{)} \\
&\geq \frac{1}{18} \cdot \gamma^2 \cdot {\beta_\ell}^2 \cdot (1-\delta_\ell)^2. \qquad (41)
\end{aligned}$$

Recall that fairAUC selects an auxiliary feature i, satisfying

$$i \in \mathrm{argmax}_{\ell \in [d']} \mathrm{AUC}_{g(t)}(S(t), Z^\ell). \tag{42}$$

Using Equations (41) and (42), we get that

$$\mathrm{AUC}_{g(t)}(S(t), Z^i) - \mathrm{AUC}_{g(t)}(X(t)) \overset{(42)}{=} \max_{\ell \in [d']} \mathrm{AUC}_{g(t)}(S(t), Z^\ell) - \mathrm{AUC}_{g(t)}(X(t)) \overset{(41)}{\geq} \max_{\ell \in [d']} \frac{1}{18} \cdot \left(\gamma \beta_\ell (1-\delta_\ell)\right)^2. \tag{43}$$

Finally, using Lemma F.9, we get that

$$\mathrm{AUC}_{g(t)}(X(t), Z^i) - \mathrm{AUC}_{g(t)}(X(t)) \overset{\text{Lemma F.9}}{\geq} \mathrm{AUC}_{g(t)}(S(t), Z^i) - \mathrm{AUC}_{g(t)}(X) \overset{(43)}{\geq} \max_{\ell \in [d']} \frac{1}{18} \cdot \left(\gamma \beta_\ell (1-\delta_\ell)\right)^2.$$

F.2.3 Proof of Lemma F.10

Proof. Proof of Lemma F.10. Equation (34) follows by using Lemma F.6 with $d = 2$. To see this, note that by Equation (15), $S(t)$ is a fixed projection of the random variable X. Since conditioned on Y and A, $X \cup Z$ is distributed according to a multivariate Gaussian distribution (Definition F.1), it follows that $X(t)$, and so $S(t)$, also has a Gaussian distribution conditioned on Y and A (see e.g., (Stirzaker 2003, Theorem 5, Section 8.4)). Now Equation (34) follows from Lemma F.6 by substituting appropriate values for the covariance matrix between $S(t)$ and Z, and the means of $S(t)$ and Z.

Equation (35) follows by expanding Equation (34). Consider the expression inside $\Phi(\sqrt{\cdot})$ in Equation (34). We have

$$\begin{bmatrix} \sigma_{0S}^2 + \sigma_{1S}^2 & \rho_{0\ell} + \rho_{1\ell} \\ (\rho_{0\ell} + \rho_{1\ell}) & \sigma_{0\ell}^2 + \sigma_{1\ell}^2 \end{bmatrix}^{-1} = \frac{1}{(\sigma_{0S}^2 + \sigma_{1S}^2) \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - (\rho_{0\ell} + \rho_{1\ell})^2} \begin{bmatrix} \sigma_{0\ell}^2 + \sigma_{1\ell}^2 & -(\rho_{0\ell} + \rho_{1\ell}) \\ -(\rho_{0\ell} + \rho_{1\ell}) & \sigma_{0S}^2 + \sigma_{1S}^2 \end{bmatrix}$$

Evaluating the rest of the expression, we have

$$\begin{aligned}
&\frac{1}{(\sigma_{0S}^2 + \sigma_{1S}^2) \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - (\rho_{0\ell} + \rho_{1\ell})^2} \cdot \begin{bmatrix} \Delta\mu_S & \Delta v_\ell \end{bmatrix} \begin{bmatrix} \sigma_{0\ell}^2 + \sigma_{1\ell}^2 & -(\rho_{0\ell} + \rho_{1\ell}) \\ -(\rho_{0\ell} + \rho_{1\ell}) & \sigma_{0S}^2 + \sigma_{1S}^2 \end{bmatrix} \begin{bmatrix} \Delta\mu_S \\ \Delta v_\ell \end{bmatrix} \\
&= \frac{1}{(\sigma_{0S}^2 + \sigma_{1S}^2) \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - (\rho_{0\ell} + \rho_{1\ell})^2} \cdot \begin{bmatrix} \Delta\mu_S \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - \Delta v_\ell \cdot (\rho_{0\ell} + \rho_{1\ell}) \\ -\Delta\mu_S \cdot (\rho_{0\ell} + \rho_{1\ell}) + \Delta v_\ell \cdot (\sigma_{0S}^2 + \sigma_{1S}^2) \end{bmatrix}^\top \begin{bmatrix} \Delta\mu_S \\ \Delta v_\ell \end{bmatrix} \\
&= \frac{1}{(\sigma_{0S}^2 + \sigma_{1S}^2) \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - (\rho_{0\ell} + \rho_{1\ell})^2} \cdot \left(\Delta\mu_S^2 \cdot \left(\sigma_{0\ell}^2 + \sigma_{1\ell}^2\right) - 2\Delta\mu_S \Delta v_\ell \cdot (\rho_{0\ell} + \rho_{1\ell}) + \Delta v_\ell^2 \cdot (\sigma_{0S}^2 + \sigma_{1S}^2)\right) \\
&\overset{(30),(33)}{=} \frac{1}{\Delta\mu_S \cdot (\sigma_{0\ell}^2 + \sigma_{1\ell}^2) - (\rho_{0\ell} + \rho_{1\ell})^2} \cdot \left(\Delta\mu_S^2 \cdot \left(\sigma_{0\ell}^2 + \sigma_{1\ell}^2\right) - 2\Delta\mu_S \cdot \Delta v_\ell \cdot (\rho_{0\ell} + \rho_{1\ell}) + \Delta v_\ell^2 \cdot \Delta\mu_S\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(\Delta\mu_S>0)}{=} \frac{1}{(\sigma_{0\ell}^2+\sigma_{1\ell}^2)-\frac{(\rho_{0\ell}+\rho_{1\ell})^2}{\Delta\mu_S}} \cdot \left(\Delta\mu_S \cdot \left(\sigma_{0\ell}^2+\sigma_{1\ell}^2\right) - 2\Delta v_\ell \cdot (\rho_{0\ell}+\rho_{1\ell}) + \Delta v_\ell^2\right) \\
&= \Delta\mu_S + \frac{1}{(\sigma_{0\ell}^2+\sigma_{1\ell}^2)-\frac{(\rho_{0\ell}+\rho_{1\ell})^2}{\Delta\mu_S}} \left((\rho_{0\ell}+\rho_{1\ell})^2 - 2\Delta v_\ell \cdot (\rho_{0\ell}+\rho_{1\ell}) + \Delta v_\ell^2\right) \\
&= \Delta\mu_S + \frac{1}{(\sigma_{0\ell}^2+\sigma_{1\ell}^2)-\frac{(\rho_{0\ell}+\rho_{1\ell})^2}{\Delta\mu_S}} \cdot \left(\Delta v_\ell - (\rho_{0\ell}+\rho_{1\ell})\right)^2 \\
&\geq \Delta\mu_S + \frac{\left(\Delta v_\ell - (\rho_{0\ell}+\rho_{1\ell})\right)^2}{\sigma_{0\ell}^2+\sigma_{1\ell}^2} && \left(\text{Using } \frac{(\rho_{0\ell}+\rho_{1\ell})^2}{\Delta\mu_S} \geq 0\right) \\
&\overset{(30)}{=} \Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu + \frac{\left(\Delta v_\ell - (\rho_{0\ell}+\rho_{1\ell})\right)^2}{\sigma_{0\ell}^2+\sigma_{1\ell}^2}. && (44)
\end{aligned}
$$

Substituting Equation (44) in Lemma F.6, and using the fact that $\Phi(\sqrt{\cdot})$ is an increasing function, we have

$$
\mathrm{AUC}_{g(t),\mathcal{D}}(S(t),Z) \geq \Phi\left(\sqrt{\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu + \frac{\left(\Delta v_\ell - (\rho_{0\ell}+\rho_{1\ell})\right)^2}{\sigma_{0\ell}^2+\sigma_{1\ell}^2}}\right). \tag{45}
$$

From Lemma F.6, we also have that

$$
\mathrm{AUC}_{g(t),\mathcal{D}}(X(t)) = \Phi\left(\sqrt{\Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu}\right).
$$

Thus, $\alpha = \Delta\mu^\top(\Sigma_0+\Sigma_1)^{-1}\Delta\mu$. Combining this with Equation (45), we get

$$
\mathrm{AUC}_{g(t)}(S(t),Z) \geq \Phi\left(\sqrt{\alpha + \frac{\left(\Delta v_\ell - (\rho_{0\ell}+\rho_{1\ell})\right)^2}{\sigma_{0\ell}^2+\sigma_{1\ell}^2}}\right).
$$

□

F.2.4 Proof of Lemma F.9

Proof. Proof. Given two vectors $v_1 \in \mathbb{R}^{d+1}$ and $v_2 \in \mathbb{R}^2$, let $\mathrm{AUC}(v_1, X, Z) \in [0,1]$ be the AUC of the linear classifier based on vector v_1 (see Definition F.2) and let $\mathrm{AUC}(v_2, S, Z) \in [0,1]$ be the AUC of the linear classifier based on vector v_2 (see Definition F.2). By definition of the AUC (Definition F.3), we have

$$
\begin{aligned}
\mathrm{AUC}(X,Z) &\coloneqq \max_{v\in\mathbb{R}^{d+1}} \mathrm{AUC}(v,X,Z), \qquad (46)\\
\mathrm{AUC}(S,Z) &\coloneqq \max_{v\in\mathbb{R}^{2}} \mathrm{AUC}(v,S,Z).
\end{aligned}
$$

Define

$$
v_2 \coloneqq \max_{v\in\mathbb{R}^2} \mathrm{AUC}(v,S,Z). \tag{47}
$$

Fix

$$
v_1 \coloneqq \begin{bmatrix} w & 0 \\ 0 & 1 \end{bmatrix} v_2 \in \mathbb{R}^{d+1}. \tag{48}
$$

Notice that the linear classifier using v_1 on X and Z, is identical to the linear classifier using v_2 on S and Z:

$$
\left\langle v_1, (X^1,\ldots,X^d,Z)\right\rangle \overset{(48)}{=} v_2^\top \begin{bmatrix} w^\top & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Z \end{bmatrix} \overset{(S\coloneqq\langle w,X\rangle)}{=} v_2^\top \begin{bmatrix} S \\ Z \end{bmatrix}. \tag{49}
$$

Using this, we have

$$
\mathrm{AUC}(X,Z) \overset{(46)}{\geq} \mathrm{AUC}(v_1,X,Z) \overset{(49)}{=} \mathrm{AUC}(v_2,S,Z) \overset{(47)}{=} \mathrm{AUC}(S,Z).
$$

□

F.3 Proof of Theorem F.5

Proof. Proof. The proof of Theorem F.5 follows from Equation (41) and Lemma F.9 in the proof of Theorem F.4. In the proof of Theorem F.4 only Equation (42) uses the fact that $g(t)$ is the disadvantaged group in iteration t. In particular, the proof of Equation (41) (which occurs before Equation (42)) does not use the fact that $g(t)$ is the disadvantaged group in iteration t. Thus, we can repeat the proof of Equation (41) by substituting $g(t)$ with $\hat{g}(t)$. This gives us that for all $\ell \in [d']\backslash Q(t)$, it holds that[7]

$$\mathrm{AUC}_{\hat{g}(t)}(S(t), Z^\ell) - \mathrm{AUC}_{\hat{g}(t)}(X(t)) \geq \frac{1}{18} \cdot (\hat{\gamma}^{(t)} \hat{\beta}_\ell^{(t)} (1 - \hat{\delta}_\ell^{(t)}))^2. \tag{50}$$

The proof of Lemma F.9 does not refer to $g(t)$. Thus, we can use it directly. This gives us that for all $\ell \in [d']\backslash Q(t)$

$$\mathrm{AUC}_{\hat{g}(t)}(X(t), Z^\ell) - \mathrm{AUC}_{\hat{g}(t)}(X(t)) \overset{\text{Lemma F.9}}{\geq} \mathrm{AUC}_{\hat{g}(t)}(S(t), Z^\ell) - \mathrm{AUC}_{\hat{g}(t)}(X) \overset{(50)}{\geq} \frac{1}{18} \cdot (\hat{\gamma}^{(t)} \hat{\beta}_\ell^{(t)} (1 - \hat{\delta}_\ell^{(t)}))^2.$$

Thus, in particular, this holds for the feature $i \in [d']\backslash Q(t)$, selected by fairAUC. $\square$

F.4 Proof of Lemma F.8

Proof. Proof of Lemma F.8.

$$\begin{aligned}
\Phi(\sqrt{\alpha + \Delta}) - \Phi(\sqrt{\alpha}) \quad &= \quad \int_{\sqrt{\alpha}}^{\sqrt{\alpha+\Delta}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\geq \quad \int_{\sqrt{\alpha}}^{\sqrt{\alpha+\Delta_0}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\qquad \text{(Using that } \Delta \geq \Delta_0 \text{ and that the RHS is an increasing function of } \Delta) \\
&\geq \quad \int_{\sqrt{\alpha}}^{\sqrt{\alpha+\Delta_0}} \frac{y}{\sqrt{\alpha+\Delta_0}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy \\
&\qquad \text{(Using the fact that for all } y \in [\sqrt{\alpha}, \sqrt{\alpha+\Delta_0}],\ \tfrac{y}{\sqrt{\alpha+\Delta_0}} \leq 1) \\
&= \quad \left. \frac{-e^{-y^2/2}}{\sqrt{\alpha+\Delta_0}} \right|_{\sqrt{\alpha}}^{\sqrt{2\pi\cdot(\alpha+\Delta_0)}} \\
&= \quad \frac{e^{-\alpha/2} \cdot \left(1 - e^{-\Delta_0/2}\right)}{\sqrt{2\pi \cdot (\alpha + \Delta_0)}} \\
&\geq \quad \frac{\gamma \cdot \left(1 - e^{-\Delta_0/2}\right)}{\sqrt{2\pi \cdot (2 \cdot \ln(1/\gamma) + \Delta_0)}} \\
&\qquad \text{(Using that for all } x \in \mathbb{R} \text{ and } \Delta_0 \geq 0,\ \tfrac{e^{-x/2}}{\sqrt{x+\Delta_0}} \text{ is decreasing in } x \text{ and that } \alpha < 2\ln(1/\gamma)) \\
&\geq \quad \frac{\gamma^2 \cdot \left(1 - e^{-\Delta_0/2}\right)}{\sqrt{2\pi \cdot (1 + \Delta_0)}} \\
&\qquad \text{(Using that for all } \gamma \in (0,1),\ \left(\tfrac{\gamma^2}{2\cdot\ln(1/\gamma)+\Delta_0}\right)^{1/2} \geq \tfrac{\gamma^2}{\sqrt{1+\Delta_0}}) \\
&\geq \quad \frac{\gamma^2 \cdot \Delta_0}{2\sqrt{2\pi \cdot (1+\Delta_0)^3}} \quad \text{(Using the fact that for all } x \in \mathbb{R},\ 1 - e^{-x/2} \geq \tfrac{x}{2(1+x)}) \\
&\geq \quad \frac{\gamma^2 \cdot \Delta_0}{6(1+\Delta_0)^{3/2}}. \qquad \text{(Using that } 2\sqrt{2\pi} \leq 6.)
\end{aligned}$$

$\square$

[7] Recall that in the proof of Theorem F.4, we dropped the superscript on $\hat{\gamma}^{(t)}$, $\hat{\beta}_\ell^{(t)}$, and $\hat{\delta}_\ell^{(t)}$. We added the superscripts back here.