Journal of Machine Learning Research 22 (2021) 1–78 Submitted 10/20; Revised 5/21; Published 10/21

# Optimized Score Transformation for Consistent Fair Classification

**Dennis Wei** DWEI@US.IBM.COM
**Karthikeyan Natesan Ramamurthy** KNATESA@US.IBM.COM
*IBM Research*
*1101 Kitchawan Road*
*Yorktown Heights, NY 10598, USA*

**Flavio P. Calmon** FLAVIO@SEAS.HARVARD.EDU
*John A. Paulson School of Engineering and Applied Sciences*
*Harvard University*
*150 Western Ave*
*Allston, MA 02134, USA*

**Editor:** Maya Gupta

## Abstract

This paper considers fair probabilistic binary classification where the outputs of primary interest are predicted probabilities, commonly referred to as scores. We formulate the problem of transforming scores to satisfy fairness constraints that are linear in conditional means of scores while minimizing a cross-entropy objective. The formulation can be applied directly to post-process classifier outputs and we also explore a pre-processing extension, thus allowing maximum freedom in selecting a classification algorithm. We derive a closed-form expression for the optimal transformed scores and a convex optimization problem for the transformation parameters. In the population limit, the transformed score function is the fairness-constrained minimizer of cross-entropy with respect to the true conditional probability of the outcome. In the finite sample setting, we propose a method called FairScoreTransformer to approach this solution using a combination of standard probabilistic classifiers and ADMM. We provide several consistency and finite-sample guarantees for FairScoreTransformer, relating to the transformation parameters and transformed score function that it obtains. Comprehensive experiments comparing to 10 existing methods show that FairScoreTransformer has advantages for score-based metrics such as Brier score and AUC while remaining competitive for binary label-based metrics such as accuracy.

**Keywords:** algorithmic fairness, machine learning fairness, probabilistic classification, post-processing

## 1. Introduction

Recent years have seen a surge of interest in the problem of *fair classification*, which is concerned with disparities in classification output or performance when conditioned on protected attributes such as race, gender, or ethnicity. Many measures of fairness have been introduced (Pedreschi et al., 2012; Dwork et al., 2012; Kamiran et al., 2013; Hardt et al., 2016; Zafar et al., 2017a; Chouldechova, 2017; Kleinberg et al., 2017; Kilbertus et al., 2017; Kusner et al., 2017; Zafar et al., 2017b; Nabi and Shpitser, 2018; Kearns et al.,

2018; Heidari et al., 2018; Chiappa, 2019) and fairness-enhancing interventions have been proposed to mitigate these disparities (Friedler et al., 2019). Roughly categorized, these interventions either (i) change data used to train a classifier (pre-processing) (Kamiran and Calders, 2012; Hajian and Domingo-Ferrer, 2013; Zemel et al., 2013; Feldman et al., 2015; Calmon et al., 2017), (ii) change a classifier's output (post-processing) (Kamiran et al., 2012; Fish et al., 2016; Hardt et al., 2016; Pleiss et al., 2017; Woodworth et al., 2017), or (iii) directly change a classification model to ensure fairness (in-processing) (Calders and Verwer, 2010; Kamishima et al., 2012; Zafar et al., 2017a,c; Dwork et al., 2018; Agarwal et al., 2018; Krasanakis et al., 2018; Donini et al., 2018; Celis et al., 2019).

This paper differs from many of the above works in placing more emphasis on probabilistic classification, in which the outputs of interest are predicted probabilities of belonging to one of the classes as opposed to binary predictions. The predicted probabilities are often referred to as *scores*. They are desirable because they indicate confidences in predictions (when well-calibrated) and provide more information for decision-making. For example, in a loan approval scenario, a score of 0.7 may indicate that a loan applicant is predicted to have a 70% chance of repaying the loan on time, given their credit history features.

Our objective is to produce probabilistic scores satisfying fairness criteria. These scores can be useful in a number of decision-making scenarios. In health risk assessment for example, the scores represent risks of developing a condition or requiring medical intervention (e.g., stroke, Lip et al., 2010, ICU admission, Zhao et al., 2020) and are the final output of interest. In other applications, the scores are an intermediate output that is passed to a subsequent decision-making stage. However, this subsequent stage may not be fully known or defined, may take additional inputs, and/or may be performed by a different party. An important example is where the decision-maker is a human (e.g., a hiring manager) who, in addition to considering a score (e.g., predicted probability of succeeding in a new job), may have to weigh other information (e.g., reports from human interviewers), and whose fairness and other decision-making properties cannot be well-controlled. In this case, it may be desirable to enforce fairness in the scores given to the human decision-maker (perhaps in addition to measures that encourage the decision-maker to be more fair). Even in the straightforward case where the scores are thresholded to produce a binary decision, exact knowledge of protected attributes may be lacking to use existing post-processing methods for fairness (Kamiran et al., 2012; Hardt et al., 2016; Pleiss et al., 2017; Yang et al., 2020). Moreover, our experimental results in Section 6 suggest that thresholding fairer scores can be competitive with fairness methods that directly target binary outputs.

We make several contributions to the subject of fair probabilistic classification. In Section 2, we propose an optimization formulation for transforming scores to satisfy fairness constraints while minimizing a cross-entropy objective. The formulation accommodates any fairness criteria that can be expressed as linear inequalities involving conditional means of scores, including variants of statistical parity (SP) (Pedreschi et al., 2012) and equalized odds (EO) (Hardt et al., 2016; Zafar et al., 2017a).

In Section 3, we study solutions to the optimization problem of fair score transformation that we have formulated. Given an input score function $r(x)$, we derive a closed-form expression for the optimal transformed scores $r'(x)$ and a convex dual optimization problem for the Lagrange multipliers that parametrize the transformation. In the population limit, the optimal input score function (i.e., the unconstrained optimum) is the conditional

distribution $p_{Y \mid X}$ of the outcome $Y$ given features $X$. In this case, the transformed scores minimize cross-entropy with respect to $p_{Y \mid X}$ while satisfying the fairness constraints.

In Section 4, we consider the finite sample setting and propose a method called FairScoreTransformer (FST) to approximate the optimal solution found in Section 3. FST takes a practical "plug-in" approach, using standard probabilistic classifiers (e.g., logistic regression) to approximate $p_{Y \mid X}$ and estimating other probabilities as needed. In particular, if protected attributes are not known at test time, FST can instead use estimates of them based on the available features. We find that the dual problem is well-suited to the alternating direction method of multipliers (ADMM) and describe an ADMM algorithm to solve it. The closed-form expression for the transformed scores and the low dimension of the dual problem (a small multiple of the number of protected groups) make FST computationally lightweight.

FST lends itself naturally to post-processing, with scores as input and fairer scores as output. To increase flexibility, we also explore a pre-processing extension of FST in which the output scores are used to re-weight the training data. The re-weighted data can then be published as an output in its own right, allowing others to train fairer models using standard algorithms that do not explicitly account for fairness. We envision therefore that FST will be particularly beneficial in situations that make post- and pre-processing attractive, as also articulated by e.g., Hajian and Domingo-Ferrer (2013); Calmon et al. (2017); Agarwal et al. (2018); Madras et al. (2018); Salimi et al. (2019): a) when it is not possible or desirable to modify an existing classifier (only post-processing is possible); b) when freedom is desired to select the most suitable classifier for an application, whether it maximizes performance or has some other desired property such as interpretability (post- and pre-processing apply); and c) when standard training algorithms are used without the additional complexity of fairness constraints or regularizers (post- and pre-processing again). In-processing meta-algorithms (Agarwal et al., 2018; Celis et al., 2019) can also support situation b) but not a) or c), while standard in-processing does not support any of a)–c). As discussed in Section 1.1 and summarized in Table 3, FST is considerably more flexible than existing post- and pre-processing methods in handling more cases.

The conference version (Wei et al., 2020) of this work focused on formulating and solving the optimization problem (Sections 2 and 3) and translating the solution into a practical procedure (Section 4). This has left a gap however between the solution in the ideal population setting ($r(x) = p_{Y \mid X}(1 \mid x)$) and the approximate result of the FST procedure. In this extended version, we address this gap by providing consistency and finite-sample guarantees. In Section 5, under suitable assumptions on the convergence of the estimated score function $\hat{r}(x)$ and other estimated probabilities, we prove that:

1. Optimal solutions to the empirical version of the dual problem solved by FST become asymptotically optimal for the population version of the dual problem. For finite sample sizes, the optimality gap is bounded with high probability (Theorem 6).

2. The plug-in solution for the transformed scores asymptotically satisfies the population fairness constraints (i.e., fairness consistency). For finite sample sizes, the degree of infeasibility is bounded with high probability (Theorem 4).

3. The plug-in solution asymptotically minimizes cross-entropy with respect to $p_{Y \mid X}$ subject to the fairness constraints (Theorem 5).

We have accordingly refined the presentation in Sections 3 and 4, for example clearly distinguishing between the empirical and population dual problems and explicitly defining the plug-in primal solution. Of note, we have clarified that the characterization of the optimal solution in Section 3 applies to any input score function $\hat{r}(x)$, not just $r(x) = p_{Y \mid X}(1 \mid x)$.

We have conducted comprehensive experiments, reported in Section 6 and Appendix C, comparing FST to 10 existing methods, a number that compares favorably to recent meta-studies (Friedler et al., 2019). On score-based metrics such as Brier score and AUC, FST achieves better fairness-utility trade-offs and hence is indeed advantageous when scores are of interest. At the same time, it remains competitive on binary label-based metrics such as accuracy.

In summary, FairScoreTransformer enables fairness-enhancing post-processing that

- is principled, optimal in the population limit, and comes with consistency and finite-sample guarantees (Sections 2, 3, and 5),
- is computationally lightweight (Section 4),
- performs favorably compared to the state-of-the-art and can handle lack of protected attributes at test time (Section 6 and Appendix C).

The organization of the paper is recapitulated below: Section 2 formulates the optimization problem of transforming scores to satisfy fairness constraints. Section 3 specifies the optimal solution to the problem in terms of a closed-form transformation and a dual optimization problem. Section 4 describes the FairScoreTransformer procedure that approximates the optimal solution given a finite sample. Section 5 provides theoretical results for the FairScoreTransformer solution. Section 6 discusses empirical evaluation of FairScoreTransformer and comparisons to existing methods. Section 7 concludes the paper.

### 1.1 Related Work

Existing post-processing methods for fairness include those from Kamiran et al. (2012); Fish et al. (2016); Hardt et al. (2016); Pleiss et al. (2017); Jiang et al. (2019); Chzhen et al. (2019); Yang et al. (2020); limitations of post-processing are studied by Woodworth et al. (2017). While these methods take predicted scores as input, most (Kamiran et al., 2012; Fish et al., 2016; Hardt et al., 2016; Chzhen et al., 2019) are designed to produce only binary output and not scores. The method of Pleiss et al. (2017) maintains calibrated probability estimates, which is a requirement that we do not enforce herein. Furthermore, Kamiran et al. (2012); Fish et al. (2016); Hardt et al. (2016); Pleiss et al. (2017); Yang et al. (2020) all assume exact knowledge of the protected attribute. Kamiran et al. (2012); Fish et al. (2016); Jiang et al. (2019) address only SP (Kamiran et al. (2012) as originally proposed), Hardt et al. (2016); Pleiss et al. (2017) address disparities in error rates, and Chzhen et al. (2019) address only equal opportunity. Our approach does not have these limitations. It produces scores as well as binary outputs, can handle estimated protected attributes, and accommodates a wider range of fairness criteria.

Pre-processing methods range from reweighing, resampling, and relabeling training data (Kamiran and Calders, 2012), to performing probability transformations on features (Feldman et al., 2015), to modifying both labels and features through optimization (Calmon et al.,

2017) or labels and protected attributes using classification rules (Hajian and Domingo-Ferrer, 2013). The above methods only address SP or the related notion of disparate impact (Feldman et al., 2015). Learning representations that are invariant to protected attributes (Zemel et al., 2013; Louizos et al., 2016; Edwards and Storkey, 2016; Xie et al., 2017; Xu et al., 2018) can also be seen as pre-processing, and recent adversarial approaches (Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018) permit control of EO as well as SP. Representation learning however does not preserve the original data domain and its semantics, while adversarial algorithms can produce unstable results and be computationally challenging.

Several works by Agarwal et al. (2018); Celis et al. (2019); Menon and Williamson (2018); Corbett-Davies et al. (2017); Jiang and Nachum (2020); Yang et al. (2020) have technical similarities to the approach herein but focus on binary outputs, with 0-1 risk (Celis et al., 2019; Agarwal et al., 2018) or cost-sensitive risk (Menon and Williamson, 2018; Corbett-Davies et al., 2017; Yang et al., 2020) as the objective function, and/or lead to in-processing algorithms (Celis et al., 2019; Agarwal et al., 2018; Cotter et al., 2019). Celis et al. (2019) come closest in also solving a fairness-constrained classification problem via the dual problem. However, Celis et al. (2019) along with Agarwal et al. (2018) propose in-processing algorithms that solve multiple instances of a subproblem whereas we solve only one instance. Celis et al. (2019) also address a larger class of fairness measures that are linear-fractional in the classifier output. Cotter et al. (2019) propose incorporating rate constraints when training predictive models in order to meet target fairness, churn, or other performance requirements. These constraints are cast in terms of indicator functions and are inherently non-convex and non-differentiable, motivating an oracle-based in-processing optimization algorithm. Unlike Cotter et al. (2019), the optimized transformation introduced here circumvents non-differentiability issues by formulating fairness constraints in terms of scores (as opposed to a sum of indicator functions). The resulting optimization is convex and solvable using standard methods.

Similar to us, Menon and Williamson (2018); Corbett-Davies et al. (2017); Yang et al. (2020) also characterize optimal fair classifiers in the population limit in which probability distributions are known; however, Menon and Williamson (2018); Corbett-Davies et al. (2017) do not propose algorithms for computing the Lagrange multipliers or thresholds that parametrize the solution. The recent work of Yang et al. (2020) provides such a characterization in a very general multi-class setting with overlapping protected groups. They propose two algorithms inspired by the Bayes-optimal fair classifier. The first is an in-processing approach that generalizes the algorithm of Agarwal et al. (2018). The second is similar to ours in also taking a plug-in post-processing approach and optimizing Lagrange multipliers. In their case, the Lagrange multipliers determine thresholds to apply to the "plugged-in" probabilistic classifier. However, both algorithms of Yang et al. (2020) return a *randomized* classifier (similar to Agarwal et al., 2018), i.e., a probability distribution over a set of classifiers, and they also assume knowledge of the protected attributes.

## 2. Problem Formulation

We represent one or more protected attributes such as gender and race by a random variable $A$ and an outcome variable by $Y$. We make the common assumption that $Y \in \{0, 1\}$ is

binary-valued. It is assumed that $A$ takes a finite number of values in a set $\mathcal{A}$, corresponding to protected groups. Let $X$ denote features (drawn from domain $\mathcal{X}$) used to predict $Y$ in a supervised classification setting. We consider two scenarios in which $X$ either includes or does not include $A$, like in other works in fair classification (e.g., Kamiran and Calders, 2012; Agarwal et al., 2018; Donini et al., 2018). While it is recognized that the former scenario can achieve better trade-offs between utility and fairness, the latter is needed in applications where disparate treatment laws and regulations forbid the explicit use of $A$. To develop our approach in this section and Section 3, we work in the population limit and make use of probability distributions involving $A$, $X$, $Y$. Section 4 discusses how these distributions are approximated using a training sample. In general, we use capital letters (e.g., $A$, $X$) to refer to random variables, and lowercase letters ($a$, $x$) to their realizations.

As stated earlier, we focus more heavily on probabilistic classification in which the output of interest is the predicted probability of being in the positive class $Y = 1$ rather than a binary prediction. The optimal probabilistic classifier is the conditional probability $r(x) \equiv p_{Y \mid X}(1 \mid x)$, which we refer to as the *population score* because it is only known in the population limit. Bayes-optimal binary classifiers can be derived from $r(x)$ by thresholding, specifically at level $c \in [0, 1]$ if $c$ and $1 - c$ are the relative costs of false positive and false negative errors. Score functions will thus play the central role in our development.

We propose a mathematical formulation and method called FairScoreTransformer (FST) that leads directly to a post-processing solution. The goal is to transform $r(x)$ into a *transformed score* $r'(x)$ that satisfies fairness conditions while minimizing the loss in optimality compared to $r(x)$. The transformed score $r'(x)$ is taken as the classification output and can be thresholded to provide a binary prediction. We elaborate on the utility and fairness measures considered in Sections 2.1 and 2.2.

We also consider a pre-processing extension of FST in which $r'(x)$ is used to transform the training data and train a new classifier, which provides the final output. For this case, we additionally define a *transformed outcome* variable $Y' \in \{0, 1\}$ and let $r'(x) = p_{Y' \mid X}(1 \mid x)$ be the conditional probability associated with it. The overall procedure consists of two steps, performed in general by two different parties: 1) The *data owner* transforms the outcome variable from $Y$ to $Y'$; 2) The *modeler* trains a classifier with $Y'$ as target variable and $X$ as input, without regard for fairness. The transformed score $r'(x)$ plays two roles in this procedure. The first is to specify the probabilistic mapping from $X$ to $Y'$ in step 1). As discussed in Section 4.4, we realize this mapping by re-weighting the training data. The second role stems from the main challenge faced by pre-processing methods, namely that the predominant fairness metrics depend on the output of the classifier trained in step 2) but this classifier is not under direct control of the pre-processing in step 1). In recognition of this challenge, we make the following assumption, also discussed by Madras et al. (2018); Salimi et al. (2019):

**Assumption 1 (pre-processing)** *The classifier trained by the modeler approximates the transformed score $r'(x)$ if it is a probabilistic classifier or a thresholded version of $r'(x)$ if it is a binary classifier.*

This assumption is satisfied for modelers who are "doing their job" in learning to predict $Y'$ from $X$ since the optimal classifier in this case is $r'(x)$ or a function thereof. Given the assumption, we will use $r'(x)$ as a surrogate for the actual classifier output. The assumption

is not satisfied if the modeler is not competent or, worse, malicious in trying to discriminate against certain protected groups.

We note that this pre-processing extension is not specific to FST and could be applied to other methods that produce a fair output score similar to $r'(x)$, for example in-processing methods that work with probabilistic classifiers.

### 2.1 Utility Measure

We propose to measure the loss in optimality, i.e., utility, between the transformed score $r'(x)$ and population score $r(x)$ using the following cross-entropy:

$$\mathbb{E}\big[-\log p_{Y' \mid X}(Y \mid X)\big] = \mathbb{E}\big[-r(X)\log r'(X) - (1-r(X))\log(1-r'(X))\big], \tag{1}$$

where the right-hand side results from expanding the expectation over $Y$ conditioned on $X$, and $p_{Y' \mid X}$ is used only as notational shorthand in the post-processing case since $Y'$ is not generated. For simplicity, we shall also use the following notation for cross-entropy:

$$H_b(p,q) \triangleq -p\log q - (1-p)\log(1-q). \tag{2}$$

The utility measure in (1) is equivalent to $\mathbb{E}\left[H_b\left(r(X), r'(X)\right)\right]$.

One way to arrive at (1) is to assume that $r'(x)$, which is the classifier output in the post-processing case and a surrogate thereof in the pre-processing extension, is evaluated against the observed outcomes $y_1, \ldots, y_n$ in a training set using the cross-entropy a.k.a. log loss. This yields the empirical version of the left-hand side of (1),

$$-\frac{1}{n}\sum_{i=1}^{n} \log p_{Y' \mid X}(y_i \mid x_i).$$

The use of log loss is well-motivated by the desire for $r'(x)$ to be close to the true conditional probability $r(x)$.

An equivalent way to motivate (1) in the pre-processing context is to measure the utility lost in transformation by the Kullback-Leibler (KL) divergence between the original and transformed joint distributions

$$D_{\mathrm{KL}}\big(p_{X,Y} \,\|\, p_{X,Y'}\big) = \mathbb{E}_{p_{X,Y}}\left[\log \frac{p_{X,Y}}{p_{X,Y'}}\right] = \mathbb{E}_{p_{X,Y}}[\log p_{Y \mid X}] - \mathbb{E}_{p_{X,Y}}[\log p_{Y' \mid X}]. \tag{3}$$

On the right-hand side, the first term depends on the data distribution but not $r'(x)$ and the second term is exactly (1).

Starting from a different premise, Jiang and Nachum (2020) proposed a similar mathematical formulation in which the arguments of the KL divergence are reversed from those in (3), i.e., the given distribution is the second argument while the distribution to be determined is the first. The form of the solution of Jiang and Nachum (2020) is therefore different from the one presented herein. The order of arguments in (3) is justified by the connection to log loss in classification discussed above. The order in (3) also agrees with the common interpretation of the first argument as a given distribution and the second argument as an approximation or deviation from the given distribution.

### 2.2 Fairness Measures

We consider fairness criteria expressible as linear inequalities involving conditional means of scores,

$$\sum_{j=1}^{J} b_{lj}\mathbb{E}\big[r'(X) \mid \mathcal{E}_{lj}\big] \leq c_l, \quad l = 1, \ldots, L, \tag{4}$$

where $\{b_{lj}\}$ and $\{c_l\}$ are real-valued coefficients and the conditioning events $\mathcal{E}_{lj}$ are defined in terms of $(A, X, Y)$ but do not depend on $r'$. Special cases of (4) correspond to the well-studied notions of statistical parity (SP) and equalized odds (EO). More precisely, we focus on the following variant of SP:

$$-\epsilon \leq \mathbb{E}[r'(X) \mid A = a] - \mathbb{E}[r'(X)] \leq \epsilon \qquad \forall a \in \mathcal{A}, \tag{5}$$

which we refer to as *mean score parity* (MSP) following Coston et al. (2019). Condition (5) corresponds to approximate mean independence of random variable $R' = r'(X)$ with respect to $A$. Similar notions can also be put in the form of (4), for example bounds on the ratio

$$1 - \epsilon \leq \frac{\mathbb{E}[r'(X) \mid A = a]}{\mathbb{E}[r'(X)]} \leq 1 + \epsilon,$$

referred to as *disparate impact* by Feldman et al. (2015), as well as *conditional statistical parity* (Kamiran et al., 2013; Corbett-Davies et al., 2017).

For EO, we add the condition $Y = y$ to the conditioning events in (5), resulting in

$$-\epsilon \leq \mathbb{E}[r'(X) \mid A = a, Y = y] - \mathbb{E}[r'(X) \mid Y = y] \leq \epsilon \qquad \forall a \in \mathcal{A},\ y \in \{0, 1\}. \tag{6}$$

For $y = 0$ (respectively $y = 1$), $\mathbb{E}[r'(X) \mid Y = y]$ is the false (true) positive rate (FPR, TPR) generalized for a probabilistic classifier, and $\mathbb{E}[r'(X) \mid A = a, Y = y]$ is the corresponding group-specific rate. Following Pleiss et al. (2017), we refer to (6) for $y = 0$ or $y = 1$ alone as approximate equality in generalized FPRs or TPRs, and to (6) for $y = 0$ and $y = 1$ together as generalized EO (GEO). The correspondences between (5), (6) and (4) are detailed in Appendix A.1.2.

The fairness measures (4) in our formulation are defined in terms of probabilistic scores. Parallel notions defined for binary predictions, i.e., by replacing $r'(X)$ with a thresholded version $\mathbf{1}(r'(X) > t)$, are more common in the literature. For example, the counterpart to (6) is (non-generalized) EO while the counterpart to (5) is called *thresholded score parity* by Coston et al. (2019). While our formulation does not optimize for these binary prediction measures, we nevertheless use them for evaluation in Section 6.

The form of (4) is inspired by but is less general than the linear conditional moment constraints of Agarwal et al. (2018), which replace $r'(X)$ in (4) by an arbitrary bounded function $g_j(A, X, Y, r'(X))$. We have restricted ourselves to (4) so that a closed-form optimal solution can be derived in Section 3. We note however that in both of the examples of Agarwal et al. (2018) and many fairness measures, $g_j(A, X, Y, r'(X)) = r'(X)$ and the additional generality is not required.

### 2.3 Optimization Problem

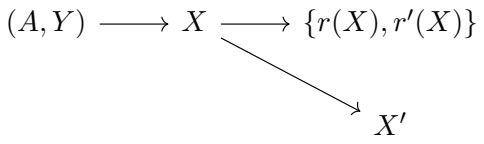
The transformed score $r'(x)$ is obtained by minimizing the cross-entropy in (1) (equivalently maximizing its negative) subject to fairness constraints (4):

$$\max_{r'} \; -\mathbb{E}\left[H_b\left(r(X), r'(X)\right)\right] \quad \text{s.t.} \quad \sum_{j=1}^{J} b_{lj}\mathbb{E}\left[r'(X) \mid \mathcal{E}_{lj}\right] \leq c_l, \quad l = 1, \ldots, L. \tag{7}$$

Section 3 characterizes the optimal solution to this problem.

### 2.4 Sufficiency of Pre-Processing Scores

In the pre-processing extension of FST, the proposed optimization (7) transforms only scores and uses them to generate a weighted data set, as described further in Section 4.4. Can a better trade-off between utility and fairness be achieved by also pre-processing features $X$, i.e., mapping each pair $(X, r(X))$ into a new $(X', r'(X))$? Note that pre-processing both scores/labels and input features is suggested by Hajian and Domingo-Ferrer (2013); Feldman et al. (2015); Calmon et al. (2017). When utility and fairness are measured according to the objective and constraints in (7), the answer is negative: a transformed feature $X'$ would not impact the constraints in (7), since they only depend on the marginals of $r'(X)$ conditioned events $\mathcal{E}_{l,j}$ given in terms of $A$ and $Y$. Moreover, a transformed feature would also not change the objective value, which only depends on $r(X)$ and $r'(X)$. In other words, a transformed score/label pair would satisfy the Markov relation:




The quantities in formulation (7) only depend on the upper branch of the above graph and, hence, are invariant to the mapping from $X$ to $X'$. Thus, for the metrics considered here, pre-processing the scores is sufficient.

## 3. Characterization of Optimal Fairness-Constrained Score

In this section, we consider a slight generalization of problem (7) in which $r(X)$ is replaced by an arbitrary score function $\hat{r}(X)$:

$$\max_{r'} \; -\mathbb{E}\left[H_b\left(\hat{r}(X), r'(X)\right)\right] \quad \text{s.t.} \quad \sum_{j=1}^{J} b_{lj}\mathbb{E}\left[r'(X) \mid \mathcal{E}_{lj}\right] \leq c_l, \quad l = 1, \ldots, L. \tag{8}$$

In later sections, $\hat{r}(X)$ will be an estimate of $r(X)$, thus justifying the hat notation.

We derive a closed-form expression for the optimal solution to problem (8) using the method of Lagrange multipliers. We then state the dual optimization problem that determines the Lagrange multipliers. These results are specialized to the cases of MSP (5) and GEO (6).

Define Lagrange multipliers $\lambda_l \geq 0$, $l = 1, \ldots, L$ for the constraints in (8), and let $\lambda \triangleq (\lambda_1, \ldots, \lambda_L)$. Then the Lagrangian function is given by

$$L(r', \lambda) = -\mathbb{E}\big[H_b\left(\hat{r}(X), r'(X)\right)\big] - \sum_{l=1}^{L}\sum_{j=1}^{J} \lambda_l b_{lj} \mathbb{E}\big[r'(X) \mid \mathcal{E}_{lj}\big] + \sum_{l=1}^{L} c_l \lambda_l. \tag{9}$$

The dual optimization problem corresponding to (7) is

$$\min_{\lambda \geq 0} \max_{r'} L(r', \lambda).$$

Note that $L(r', \lambda)$ is a strictly concave function of $r'$ and the fairness constraints in (8) are affine functions of $r'$. Consequently, as long as the constraints in (8) are feasible, the optimal transformed score $r^*$ can be found by maximizing $L(r', \lambda)$ with respect to $r'$, resulting in an optimal solution $r^*$ that is a function of $\lambda$, and then minimizing $L(r^*, \lambda)$ with respect to $\lambda$ (Boyd and Vandenberghe, 2004, Section 5.5.5). Substituting the optimal $\lambda^*$ into the solution for $r^*$ found in the first step then yields the optimal transformed score. Note that this procedure would not necessarily be correct if a linear objective function were considered (e.g., 0-1 loss in Celis et al., 2019) due to lack of strict concavity. The next proposition states the general form of the solution to the inner maximization of $L(r', \lambda)$ above. Its proof is in Appendix A.1.1.

**Proposition 1** *Let $L(r', \lambda)$ be as given in (9). Then for fixed $\lambda$, $r^*(\lambda) = \arg\max_{r'} L(r', \lambda)$ is given by*

$$r^*\big(\mu(x); \hat{r}(x)\big) = \begin{cases} \dfrac{1 + \mu(x) - \sqrt{(1+\mu(x))^2 - 4\hat{r}(x)\mu(x)}}{2\mu(x)}, & \mu(x) \neq 0 \\ \hat{r}(x), & \mu(x) = 0, \end{cases} \tag{10}$$

*where*

$$\mu(x) \triangleq \sum_{l=1}^{L}\sum_{j=1}^{J} \lambda_l b_{lj} \frac{\Pr(\mathcal{E}_{lj} \mid X = x)}{\Pr(\mathcal{E}_{lj})}. \tag{11}$$

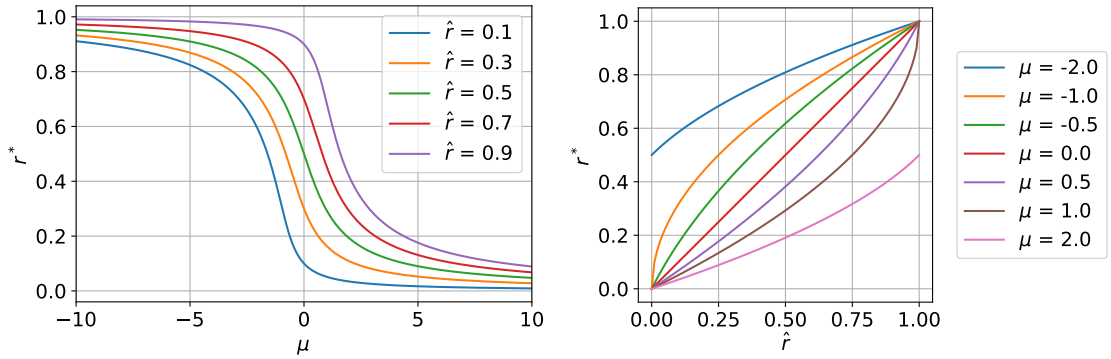
We can interpret the optimal primal solution (10) as a prescription for *score transformation* controlled by $\mu(x)$, which is in turn a linear function of $\lambda$. When $\mu(x) = 0$, the score is unchanged from the input $\hat{r}(x)$, and as $\mu(x)$ increases or decreases away from zero, the score $r^*(\mu(x); \hat{r}(x))$ decreases or increases smoothly from $\hat{r}(x)$, as seen in Figure 1a. Figure 1b shows that the transformed score $r^*$ has a rank-preserving property stated in Lemma 2.

**Lemma 2** *The transformed score $r^*(\mu; \hat{r})$ is monotonically increasing in $\hat{r}$ for fixed $\mu$, i.e., if $r_1 < r_2$ then $r^*(\mu; r_1) < r^*(\mu; r_2)$.*

**Proof** This is confirmed analytically by a positive partial derivative:

$$\frac{\partial r^*(\mu; \hat{r})}{\partial \hat{r}} = \frac{1}{\sqrt{(1+\mu)^2 - 4\hat{r}\mu}} > 0.$$

■




(a) $r^*$ vs. $\mu$ for fixed values of $\hat{r}$.

(b) $r^*$ vs. $\hat{r}$ for fixed values of $\mu$.

Figure 1: Optimal transformed score $r^*(\mu;\hat{r})$ (Equation 10) as a function of $\mu$ and $\hat{r}$.

It is shown in Appendix A.1.1 that the result of substituting the optimal primal solution (10) into the first two terms of the Lagrangian (9) is the expectation of the function

$$g\big(\mu(x);\hat{r}(x)\big) \triangleq -H_b\Big(\hat{r}(x), r^*\left(\mu(x);\hat{r}(x)\right)\Big) - \mu(x) r^*\big(\mu(x);\hat{r}(x)\big). \tag{12}$$

The dual problem is therefore

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X);\hat{r}(X)\big)\right] + \sum_{l=1}^{L} c_l \lambda_l \\ \text{s.t.} \quad & \mu(X) = \sum_{l=1}^{L}\sum_{j=1}^{J} \lambda_l b_{lj} \frac{\Pr(\mathcal{E}_{lj} \mid X)}{\Pr(\mathcal{E}_{lj})}, \qquad \lambda \geq 0. \end{aligned} \tag{13}$$

The solution to the above minimization provides the values of $\lambda^*$ for the optimal transformed score (10). Like all Lagrangian duals, (13) is a convex optimization (although it is no longer apparent from Equation 13 that this is the case). Furthermore, (13) is typically low-dimensional in cases where the number of dual variables $L$ is a small multiple of the number of protected groups $|\mathcal{A}|$.

We now specialize and simplify (13) to MSP (5) and GEO (6) fairness constraints. The following proposition follows from the correspondences between (5), (6) and (4) and is proved in Appendix A.1.2.

**Proposition 3** *Under the MSP constraint* (5)*, the dual optimization* (13) *reduces to*

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X);\hat{r}(X)\big)\right] + \epsilon\|\lambda\|_1 \\ \text{s.t.} \quad & \mu(X) = \sum_{a\in\mathcal{A}} \lambda_a \left(\frac{p_{A\,|\,X}(a \mid X)}{p_A(a)} - 1\right). \end{aligned} \tag{14}$$

*For the GEO constraint* (6)*,* (13) *reduces to*

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X);\hat{r}(X)\big)\right] + \epsilon\|\lambda\|_1, \\ \text{s.t.} \quad & \mu(X) = \sum_{y\in\{0,1\}} \frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)} \sum_{a\in\mathcal{A}} \lambda_{a,y} \left(\frac{p_{A\,|\,X,Y}(a\,|\,X,y)}{p_{A\,|\,Y}(a\,|\,y)} - 1\right). \end{aligned} \tag{15}$$

*In the case* $\hat{r}(X) = r(X)$*, we refer to* (14)*,* (15) *as the* population dual *problem.*

In (14), (15), there is no longer a non-negativity constraint on $\lambda$ but instead an $\ell_1$ norm, and the problem dimension is only $|\mathcal{A}|$ in (14) and $2|\mathcal{A}|$ in (15). Moreover, both dual formulations are well-suited for decomposition using the alternating direction method of multipliers (ADMM), as discussed further in Section 4.2.

In the case where the features $X$ include the protected attribute $A$, we have $p_{A\,|\,X}(a\,|\,X) = p_{A\,|\,X,Y}(a\,|\,X,y) = \mathbf{1}(a = A)$, where $A$ is the component of $X$ that is given. The constraints in (14) and (15) then simplify to

$$\mu(X) = \frac{\lambda_A}{p_A(A)} - \sum_{a\in\mathcal{A}} \lambda_a, \tag{16}$$

$$\mu(X) = \sum_{y\in\{0,1\}} \frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)} \left(\frac{\lambda_{A,y}}{p_{A\,|\,Y}(A\,|\,y)} - \sum_{a\in\mathcal{A}} \lambda_{a,y}\right) \tag{17}$$

respectively. Interestingly, the only difference between the cases of including or excluding $A$ is that in the latter, the constraints in (14), (15) indicate that $A$ should be inferred from the available features $X$ and possibly $Y$, whereas in the former, $A$ can be used directly.

### 3.1 Comparison to Optimal Fair Binary Classifiers

As discussed in Section 1.1, optimal fair classifiers have been characterized by Menon and Williamson (2018); Corbett-Davies et al. (2017); Yang et al. (2020) in the case of binary outputs and cost-sensitive risk. While the score transformation discussed herein is optimized for different, score-based measures of utility and fairness, it is still of interest to compare the result of thresholding the transformed score to these optimal fair binary-output classifiers.

We focus on Menon and Williamson (2018), who give the most concrete expressions for optimal classifiers compared to Corbett-Davies et al. (2017); Yang et al. (2020). In accordance with Menon and Williamson (2018), we consider the population limit, e.g., $\hat{r}(X) = r(X)$, use a binary protected attribute, $\mathcal{A} = \{0, 1\}$, and consider the fairness measures statistical parity (SP) and equal opportunity (EOpp) with respect to $A$. To be closer to fairness constraints (5), (6), we consider the "mean difference" (MD) measure of Menon and Williamson (2018, Equation 6). In the SP-MD case, Menon and Williamson (2018) minimize the following cost-sensitive risk (Problem 3.2 therein):

$$(1-c)\Pr(\hat{Y}=0|Y=1)+c\Pr(\hat{Y}=1|Y=0)-\lambda\left(\Pr(\hat{Y}=1\,|\,A=0)-\Pr(\hat{Y}=1\,|\,A=1)\right), \tag{18}$$

| fairness criterion | $A$ known | Menon and Williamson | $h(X)$ from optimal fair score |
|---|---|---|---|
| SP | no | $r(X) - \lambda(\bar{\eta}(X) - 1/2)$ | $r(X) - c(1-c)\left(\lambda_0\theta_0(X) + \lambda_1\theta_1(X)\right),$ $\theta_a(X) = \frac{(1-\bar{\eta}(X))^{1-a}\bar{\eta}(X)^a}{p_A(a)} - 1, \;\; a = 0, 1$ |
| | yes | $r(X) + (-1)^A(1/2)\lambda$ | $r(X) + (-1)^A c(1-c)p_A(1-A)\tilde{\lambda},$ $\tilde{\lambda} = \frac{\lambda_1}{p_A(1)} - \frac{\lambda_0}{p_A(0)}$ |
| EOpp | no | $\left(1 - \frac{\lambda}{p_{Y}(1)}(\bar{\eta}(X) - 1/2)\right) r(X)$ | $\left(1 - \frac{c(1-c)}{p_Y(1)}\left(\lambda_0\theta_0(X) + \lambda_1\theta_1(X)\right)\right) r(X),$ $\theta_a(X) = \left(\frac{(1-\bar{\eta}(X))^{1-a}\bar{\eta}(X)^a}{p_{A \mid Y}(a \mid 1)} - 1\right)$ |
| | yes | $\left(1 + (-1)^A \frac{1}{2p_Y(1)}\lambda\right) r(X)$ | $\left(1 + (-1)^A \frac{c(1-c)p_{A \mid Y}(1-A \mid 1)}{p_Y(1)}\tilde{\lambda}\right) r(X),$ $\tilde{\lambda} = \left(\frac{\lambda_1}{p_{A \mid Y}(1 \mid 1)} - \frac{\lambda_0}{p_{A \mid Y}(0 \mid 1)}\right)$ |

Table 1: Comparison with optimal fair binary classifiers of Menon and Williamson (2018). All binary classifiers are of the form $\mathbf{1}(h(X) > c)$ where $h(X)$ is given below.

where the first and second terms are the FNR and FPR, weighted by $1 - c$ and $c$, and the last two terms are the difference between positive prediction rates. For EOpp, the last two terms are additionally conditioned on $Y = 1$.

Table 1 summarizes the expressions for fair binary classifiers, which are all of the form $\mathbf{1}(h(X) > c)$ where $h(X)$ is given in the table. For the rightmost column, we assume that the optimal transformed score $r^*(\mu(X); r(X))$ is thresholded at the cost-sensitive threshold $c$ to obtain a binary prediction. Derivations of all expressions are given in Appendix A.1.3. We use the notation $\bar{\eta}(X)$ of Menon and Williamson (2018) for the conditional probability of $A$ given $X$, where $\bar{\eta}(X) = p_{A \mid X}(1 \mid X)$ in the SP case and $\bar{\eta}(X) = p_{A \mid X,Y}(1 \mid X, 1)$ for EOpp.

Overall, while the expressions from Menon and Williamson (2018) and from thresholding the optimal transformed score are different, they do have notable similarities. The four cases in Table 1 are discussed further below, where "$A$ known" means that $A$ is included in $X$ or is perfectly predicted by $X$.

- **SP, $A$ not known:** The two $h(X)$ expressions are similar in that an affine function of $\bar{\eta}(X)$ is added to the original score $r(X)$. In the case of Menon and Williamson (2018), the affine function is proportional to the trade-off parameter $\lambda$, whereas for the thresholded optimal fair score, the affine function is proportional to $c(1-c)$ and also depends on $p_A(a)$. The parameters $\lambda_0$, $\lambda_1$ are chosen to optimize the dual objective (14).

- **SP, $A$ known:** In this case, the similarity between the two expressions becomes more apparent. On the right-hand side, the quantity $\tilde{\lambda} = \lambda_1/p_A(1) - \lambda_0/p_A(0)$ plays the role of $\lambda$ on the left side, and the main difference is the scaling by the factor $p_A(1-A)$ ($p_A(1)$ for $A = 0$ and $p_A(0)$ for $A = 1$), in addition to the factor $c(1-c)$. While these differences are likely due to optimizing for different criteria, the overall

similarity between the two formulas is noteworthy. Indeed, if $p_A(0) = p_A(1)$, then the expressions coincide after defining $\lambda_0$, $\lambda_1$ appropriately.

- **EOpp, $A$ not known:** The two expressions again share similarities: now the modification to $r(X)$ is multiplicative, the multiplicative factor is affine in $\bar{\eta}(X)$, and $p_Y(1)$ appears in the denominator on both sides.
- **EOpp, $A$ known:** As in the SP, $A$ known case, the quantity $\tilde{\lambda} = \lambda_1/p_{A\,|\,Y}(1\,|\,1) - \lambda_0/p_{A\,|\,Y}(0\,|\,1)$ plays the role of $\lambda$ and the scale factors are $p_{A\,|\,Y}(1\,|\,1)$ for $A = 0$ and $p_{A\,|\,Y}(0\,|\,1)$ for $A = 1$. These are the analogues of the quantities in the SP, $A$ known case, now conditioned on $Y = 1$. Again if $p_{A\,|\,Y}(0\,|\,1) = p_{A\,|\,Y}(1\,|\,1)$, then the two expressions are equivalent.

## 4. Proposed FairScoreTransformer Procedure

We now consider the finite sample setting in which the probability distributions of $A, X, Y$ are not known and we have instead a training set $\mathcal{D}_n \triangleq \{(a_i, x_i, y_i), i = 1, \ldots, n\}$. This section presents the proposed FairScoreTransformer (FST) procedure that approximates the optimal fairness-constrained score in Section 3. We focus on the cases of MSP and GEO. The procedure consists of the following steps:

1. Estimate the population score and other probabilities required to define the dual problem (14) or (15).
2. Solve the dual problem to obtain dual variables $\hat{\lambda}$ (the "fit" step).
3. Transform scores using (11) and (10) ("transform" step).
4. For the pre-processing extension of FST, modify the training data.
5. For binary-valued predictions, binarize scores.

The following subsections elaborate on steps 1–4. Step 5 is done simply by selecting a threshold $t \in [0, 1]$ to maximize accuracy on the training set.

### 4.1 Estimation of Original Score and Other Probabilities

In some post-processing applications, estimates $\hat{r}(x)$ of the population scores $r(x)$ may already be provided by an existing base classifier. If no suitable base classifier exists, any probabilistic classification algorithm may be used to estimate $r(x)$. We experiment with logistic regression and gradient boosting machines in Section 6. We naturally recommend selecting a model and any hyperparameter values to maximize performance in this regard, i.e., to yield accurate and calibrated probabilities. This can be done through cross-validation on the training set using an appropriate metric such as Brier score (Hernández-Orallo et al., 2012).

In the case where $A$ is one of the features in $X$, the other probabilities required are $p_A(a)$ for MSP (16) and $p_Y(y)$, $p_{A\,|\,Y}(a\,|\,y)$ for GEO (17) ($p_{Y\,|\,X}(y\,|\,x)$ is already estimated by $\hat{r}(x)$ and $p_{A\,|\,X}$, $p_{A\,|\,X,Y}$ are delta functions). Since $Y$ is binary and $|\mathcal{A}|$ is typically small, it suffices to use the empirical estimates of these probabilities. If $A$ is not included in $X$,

then it is also necessary to estimate it using $p_{A\,|\,X}(a\,|\,X)$ for MSP (14) and $p_{A\,|\,X,Y}(a\,|\,X,y)$ for GEO (15). Again, any probabilistic classification algorithm can be used, provided that it can handle more than two classes if $|\mathcal{A}| > 2$.

We highlight that FST translates the effort of ensuring fair classification into training well-calibrated models for predicting $Y$ and, if necessary, $A$. This echoes the plug-in approach advocated by Menon and Williamson (2018); Chzhen et al. (2019).

### 4.2 ADMM for Optimizing Dual Variables

In the finite sample case, we solve an empirical version of the dual problem in Proposition 3. We write $\mu(x) = \lambda^T \mathbf{f}(x)$, where $\mathbf{f} : \mathcal{X} \to \mathbb{R}^L$ is defined by the expression for $\mu(x)$ in (14) or (15) (explicit definitions for $\mathbf{f}$ are given in Equations 24, 25 for the case where $A$ is known exactly), and $L$ is the dimension of $\lambda$. Let $\hat{r}(x)$ denote the estimate of $r(x)$ obtained in Section 4.1, and $\hat{\mathbf{f}}(x)$ be an empirical version of $\mathbf{f}(x)$ in which all probabilities (e.g., $p_A(a)$ for MSP in Equations 14, 16) are replaced by their estimates, again as discussed in Section 4.1. With these definitions, both optimizations in Proposition 3 have the general form

$$\min_{\lambda \in \mathbb{R}^L} \quad \frac{1}{n} \sum_{i=1}^{n} g\big(\mu(x_i); \hat{r}(x_i)\big) + \epsilon \|\lambda\|_1 \quad \text{s.t.} \quad \mu(x_i) = \lambda^T \hat{\mathbf{f}}(x_i), \quad i = 1, \ldots, n, \tag{19}$$

where the expectation in the objective has also been approximated by the average over the training data set.

Formulation (19) is well-suited for ADMM because the objective function is separable between $\mu(x)$ and $\lambda$, which are linearly related through the constraint. We present one ADMM decomposition here and alternatives in Appendix B.2. Under the first decomposition, application of the scaled ADMM algorithm (Boyd et al., 2011, Section 3.1.1) to (19) yields the following three steps in each iteration $k = 0, 1, \ldots$:

$$\mu^{(k+1)}(x_i) = \arg\min_{\mu} \; \frac{1}{n} g\big(\mu; \hat{r}(x_i)\big) + \frac{\rho}{2} \left(\mu - (\lambda^{(k)})^T \hat{\mathbf{f}}(x_i) + c^{(k)}(x_i)\right)^2 \quad \forall i = 1, \ldots, n \tag{20a}$$

$$\lambda^{(k+1)} = \arg\min_{\lambda} \; \epsilon \|\lambda\|_1 + \frac{\rho}{2} \sum_{i=1}^{n} \left(\mu^{(k+1)}(x_i) - \lambda^T \hat{\mathbf{f}}(x_i) + c^{(k)}(x_i)\right)^2 \tag{20b}$$

$$c^{(k+1)}(x_i) = c^{(k)}(x_i) + \mu^{(k+1)}(x_i) - \left(\lambda^{(k+1)}\right)^T \hat{\mathbf{f}}(x_i) \quad \forall i = 1, \ldots, n. \tag{20c}$$

Here $c^{(k)}(x_i)$ are Lagrange multipliers for the $n$ equality constraints in (19).

The first update (20a) can be computed in parallel for each sample $x_i$ in the data set. Given an $x_i$, finding $\mu(x_i)$ is a single-parameter optimization where the objective possesses closed-form expressions for its derivatives. For simplicity of notation, let $\hat{r}_i \triangleq \hat{r}(x_i)$, $b_i \triangleq (\lambda^{(k)})^T \hat{\mathbf{f}}(x_i) - c^{(k)}(x_i)$, and

$$\mathsf{obj}(\mu) \triangleq \frac{1}{n} g\big(\mu; \hat{r}_i\big) + \frac{\rho}{2} \left(\mu - b_i\right)^2 .$$

The first two derivatives of $\mathsf{obj}(\mu)$ are

$$\frac{\partial \mathsf{obj}(\mu)}{\partial \mu} = -\frac{r^*(\mu;\hat{r}_i)}{n} + \rho(\mu - b_i), \qquad \frac{\partial^2 \mathsf{obj}(\mu)}{\partial \mu^2} = \begin{cases} \frac{1}{2n\mu^2}\left(1 - \frac{1+\mu(1-2\hat{r}_i)}{\sqrt{(1+\mu)^2-4\hat{r}_i\mu}}\right) + \rho, & \mu \neq 0, \\ \frac{r(1-r)}{n} + \rho, & \mu = 0, \end{cases}$$

using (90), (91) in Appendix B.2 for the derivatives of $g(\mu;\hat{r}_i)$. It can be confirmed from the second derivative that $\mathsf{obj}(\mu)$ is convex (expected since the dual problem is convex) so that the first-order condition $\partial \mathsf{obj}(\mu)/\partial\mu = 0$ is necessary and sufficient for optimality. In Appendix B.1, we show that this condition leads to a cubic equation with a closed-form solution.

The second update (20b) reduces to an $\ell_1$-penalized quadratic minimization over (at most) $2|A|$ variables. Specifically,

$$\lambda^{(k+1)} = \arg\min_{\lambda} \ \epsilon\|\lambda\|_1 + \lambda^T \mathbf{v} + \lambda^T \mathbf{F} \lambda, \tag{21}$$

where

$$\mathbf{v} \triangleq -\rho \sum_{i=1}^{n} \hat{\mathbf{f}}(x_i)\left(\mu^{(k+1)}(x_i) + c^{(k)}(x_i)\right), \qquad \mathbf{F} \triangleq \frac{\rho}{2}\sum_{i=1}^{n} \hat{\mathbf{f}}(x_i)\hat{\mathbf{f}}(x_i)^T.$$

The ADMM approach thus handles the non-smooth $\ell_1$ term in the objective (19) by solving $\ell_1$-penalized quadratic subproblems (21), for which many solvers exist. Moreover, the values of $\mathbf{v}$ and $\mathbf{F}$ above can be pre-computed prior to solving (21). In fact, $\mathbf{F}$ can be computed once at the start of the iterations. The ensuing minimization only involves $|\mathcal{A}|$ variables under the MSP constraint (5), and $2|\mathcal{A}|$ variables under the GEO constraint (6).

From (20a)–(20c), it is seen that the computational complexity of each ADMM iteration scales linearly with $n$. We have fixed the ADMM penalty parameter $\rho = 1$ and have not attempted to tune it for faster convergence.

### 4.3 Score Transformation

Let $\hat{\lambda}$ denote an optimal solution to the empirical dual problem (19). We propose using a plug-in solution for the transformed score $r'(x)$, obtained by substituting finite-sample estimates into formula (10) for $r^*$, namely $r(x) = \hat{r}(x)$ and $\mu(x) = \hat{\lambda}^T\hat{\mathbf{f}}(x)$:

$$r'(x) = r^*\big(\hat{\lambda}^T\hat{\mathbf{f}}(x); \hat{r}(x)\big). \tag{22}$$

Sections 5.4.2 and 5.4.3 discuss the consistency properties of this plug-in solution.

### 4.4 Additional Steps for Pre-Processing

In the pre-processing extension of FST, the transformed score $r'(x)$ is used to generate samples of a transformed outcome $Y'$. Since $r'(x) = p_{Y'\,|\,X}(1\,|\,x)$ is a probabilistic mapping, we propose generating a *weighted* data set $\mathcal{D}' = \{(x_i, y_i', w_i)\}$ with weights $w_i$ that reflect the conditional distribution $p_{Y'\,|\,X}$. Specifically, $\mathcal{D}' = \mathcal{D}_0' \cup \mathcal{D}_1'$ with $\mathcal{D}_0' = \{(x_i, 0, 1 - r'(x_i)), i = 1, \ldots, n\}$ and $\mathcal{D}_1' = \{(x_i, 1, r'(x_i)), i = 1, \ldots, n\}$. With these weights, $Y'$ follows the conditional distribution given by $r'(x)$, and $\mathcal{D}'$ is twice the size of the original data set. The data owner passes the transformed data set $\mathcal{D}'$ to the modeler, who uses it to train a classifier for $Y'$ given $X$ without fairness constraints. Per Assumption 1, the output of this new classifier is expected to approximate $r'(x)$.

## 5. Consistency and Finite-Sample Guarantees for FairScoreTransformer

In this section, we present results guaranteeing the consistency of the FST procedure of Section 4, again focusing on the cases of MSP and GEO. For two of the three theorems presented, finite-sample bounds are also provided. We consider in particular steps 2 and 3 of the procedure and make the following statements respectively:

1. Optimal solutions to the empirical dual problem (19) become asymptotically optimal for the population dual problem (Equations 14 or 15 with $\hat{r}(X) = r(X)$) as the sample size $n \to \infty$ and the estimates $\hat{r}(x)$, $\hat{\mathbf{f}}(x)$ converge to their respective true quantities. For finite sample sizes, the optimality gap is bounded with high probability.

2. The finite-sample plug-in solution (22) for the transformed score $r'(x)$ becomes asymptotically feasible and optimal for the population primal problem (7), again as $n \to \infty$ and $\hat{r}(x)$, $\hat{\mathbf{f}}(x)$ converge. For finite sample sizes, the degree of infeasibility is bounded with high probability.

Asymptotic feasibility in statement 2 may also be referred to as *fairness consistency*, in that score functions that satisfy the fairness constraints on the training data also asymptotically satisfy them on the population.

We first summarize the assumptions that are made before formally stating the results. This is followed by more detailed discussion of the assumptions, their basic implications, and outlines of the proofs. Proofs of lemmas are deferred to Appendix A.

### 5.1 Assumptions

To simplify the proofs, we assume in this section that $A$ is available at test time, as stated below for easy reference:

**Assumption 2** *The protected attributes $A$ are known at test time.*

We make the assumption that the probabilities $p_A(a)$ (MSP case) and $p_{A,Y}(a, y)$ (GEO case) together with their estimates are bounded away from zero.

**Assumption 3** *For the MSP case, $p_A(a)$ and its estimate $\hat{p}_A(a)$ are bounded away from zero, i.e., $p_A(a) \geq \eta$ and $\hat{p}_A(a) \geq \eta$ for all $a \in \mathcal{A}$ and some $\eta > 0$. For the GEO case, $p_{A,Y}(a, y) \geq \eta$ and $\hat{p}_{A,Y}(a, y) \geq \eta$ for all $a \in \mathcal{A}$, $y \in \{0, 1\}$, and some $\eta > 0$.*

To ensure consistency of FST, we naturally assume that $\hat{r}(X)$ is a consistent estimator of the population score $r(X)$. More specifically, we assume for theoretical purposes that $\hat{r}(X)$ is estimated from a data set of size $m$ that is independent of the data set of size $n$ used to approximate the expectation in (19) (this might be obtained by splitting a larger data set into subsets of size $m$ and $n$.) The finite-sample bounds in the assumptions and theorems below are thus stated in terms of $m$. Different definitions of consistency suffice to prove different results. For the first definition, we view $\hat{r}(X)$ and $r(X)$ as random variables over $[0, 1]$ induced by $X$ and define $D_{\mathrm{TV}}(R_1, R_2)$ to be the total variation distance between two such random variables,

$$D_{\mathrm{TV}}(R_1, R_2) = \sup_{\mathcal{R} \subset [0,1]} |\Pr(R_1 \in \mathcal{R}) - \Pr(R_2 \in \mathcal{R})|.$$

**Assumption 4** *There exists a bound $E_{\mathrm{TV}}(m,\delta)$ as a function of $m$ and $\delta \in (0,1]$ such that*

1. *With probability at least $1-\delta$,*

$$\sum_{a\in\mathcal{A}} p_A(a) D_{\mathrm{TV}}\big(\hat{r}(X) \mid A=a, r(X) \mid A=a\big) \leq E_{\mathrm{TV}}(m,\delta);$$

2. *$E_{\mathrm{TV}}(m,\delta)$ is decreasing in $m$ for fixed $\delta$ (and decreases to zero as $m\to\infty$);*

3. *$E_{\mathrm{TV}}(m,\delta)$ is increasing in $1/\delta$ for fixed $m$.*

The second definition involves convergence of $\hat{r}(X)$ to $r(X)$ in $L_1$ norm:

**Assumption 5** *There exists a bound $E_{L_1}(m,\delta)$ as a function of $m$ and $\delta \in (0,1]$ such that*

1. *With probability at least $1-\delta$,*

$$\mathbb{E}\left[|\hat{r}(X) - r(X)|\right] \leq E_{L_1}(m,\delta);$$

2. *$E_{L_1}(m,\delta)$ is decreasing in $m$ for fixed $\delta$ (and decreases to zero as $m\to\infty$);*

3. *$E_{L_1}(m,\delta)$ is increasing in $1/\delta$ for fixed $m$.*

The third definition requires $\hat{r}(X)$ to converge to $r(X)$ in terms of the expectation of a Kullback-Leibler (KL) divergence. Define

$$D_{\mathrm{KL}}(p \,\|\, q) = p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \tag{23}$$

to be the KL divergence between Bernoulli random variables with parameters $p$ and $q$. The following assumption is stated only in terms of convergence in probability (as $m\to\infty$) as we do not make use of finite-sample bounds.

**Assumption 6** *The estimate $\hat{r}(X)$ converges to the population score $r(X)$ such that* $\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X))\right] \xrightarrow{p} 0$.

Note that the expectations in Assumptions 5 and 6 are with respect to $X$.

Lastly, we use the following assumption to show that it is sufficient to consider a bounded feasible set for the dual problem.

**Assumption 7** *The fairness constraint parameter $\epsilon > 0$.*

We also require a technical assumption to prove one of the lemmas, which we discuss in Section 5.4.3.

### 5.2 Results

Property 2 stated at the beginning of Section 5 is of primary importance as it pertains to the overall plug-in solution (22) for the primal problem (7). The first theorem below addresses the degree to which the plug-in solution satisfies the population fairness constraints.

**Theorem 4** *In the MSP case, under Assumptions 2, 3, 7, with probability at least* $1-\delta_1-\delta_2$ *and* $m > (2/\eta)\log(2L/\delta_1)$*, the finite-sample plug-in solution* $r'(x)$ *in* (22) *satisfies*

$$\begin{aligned}\left|\mathbb{E}[r'(X) \mid A=a] - \mathbb{E}[r'(X)]\right| &\leq \epsilon + \frac{\sqrt{2\log(2L/\delta_1)}}{\sqrt{mp_A(a)} - \sqrt{2\log(2L/\delta_1)}} \\ &+ \left(\frac{1}{\eta}-1\right)\left(\frac{4\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}} + \sqrt{\frac{2\log(2L/\delta_2)}{n}} + \frac{2}{\sqrt{n}}\right) \quad \forall\, a \in \mathcal{A},\end{aligned}$$

*where the terms after* $\epsilon$ *represent the excess with respect to the population fairness constraint* (5)*. In the GEO case, under Assumptions 2, 3, 5, 7, with probability at least* $1-\delta_1-\delta_2-\delta_3$ *and* $m > (2/\eta)\log(2(L+2)/\delta_1)$*, the plug-in solution satisfies*

$$\begin{aligned}&\left|\mathbb{E}[r'(X) \mid A=a, Y=y] - \mathbb{E}[r'(X) \mid Y=y]\right| \\ &\leq \epsilon + \frac{\sqrt{2\log(2(L+2)/\delta_1)}}{\sqrt{mp_{A,Y}(a,y)} - \sqrt{2\log(2(L+2)/\delta_1)}} + \frac{\sqrt{2\log(2(L+2)/\delta_1)}}{\sqrt{mp_Y(y)} - \sqrt{2\log(2(L+2)/\delta_1)}} \\ &\quad + \left(\frac{1}{\eta}-1\right)\left(\frac{4\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}} + \sqrt{\frac{2\log(2L/\delta_2)}{n}} + \frac{2}{\sqrt{n}}\right) \\ &\quad + \left(\frac{1}{\eta}-1\right) E_{L_1}(m, \delta_3) \quad \forall\, a \in \mathcal{A},\ y \in \{0,1\},\end{aligned}$$

*where the terms after* $\epsilon$ *are the excess with respect to constraint* (6)*.*

The next theorem asserts the asymptotic optimality of the plug-in primal solution.

**Theorem 5** *Under Assumptions 2, 3, 4, 6, 7, 9 and as* $n \to \infty$*,*

$$-\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); \hat{r}(X)\right)\right)\right] + \mathbb{E}\left[H_b\left(r(X), r^*\left(\lambda^{*T}\mathbf{f}(A, r(X)); r(X)\right)\right)\right] \xrightarrow{p} 0.$$

*The first term is the population primal objective evaluated at the plug-in solution while the second term is the optimal objective value.*

Unlike Theorems 4 and 6, Theorem 5 does not provide finite-sample guarantees. We discuss reasons for not doing so in Section 5.4.3.

Property 1 (beginning of Section 5) pertains to the near optimality of empirical dual solutions. It is used to prove Theorem 5 and may also be of independent interest. Let $J(\lambda)$ and $\hat{J}(\lambda)$ denote the objective functions in the population dual (14), (15) and empirical dual (19) respectively.

**Theorem 6** *Let $\hat{\lambda} \in \arg\min \hat{J}(\lambda)$ and $\lambda^* \in \arg\min J(\lambda)$ be optimal solutions to the empirical dual problem* (19) *and population dual problem* (14), (15) *respectively. Under Assumptions 2, 3, 4, 7, with probability at least $1 - \delta_1 - \delta_2 - \delta_3$ and $m > (2/\eta)\log(2(L+2)/\delta_3)$, we have*

$$J(\hat{\lambda}) - J(\lambda^*) \leq 2\log(2)\left(1 + \frac{1}{\epsilon}\left(\frac{1}{\eta} - 1\right)\right)\left(4\sqrt{\frac{2\log(2L)}{n}} + \sqrt{\frac{2\log(2/\delta_1)}{n}} + E_{\mathrm{TV}}(m, \delta_2)\right) + \Delta,$$

*where in the MSP case,*

$$\Delta = \frac{2\log 2}{\epsilon} \sum_{a \in \mathcal{A}} \frac{\sqrt{2\log(2L/\delta_3)}}{\sqrt{m p_A(a)} - \sqrt{2\log(2L/\delta_3)}},$$

*in the GEO case,*

$$\Delta = \frac{2\log 2}{\epsilon}\left(\sum_{y \in \{0,1\}} \sum_{a \in \mathcal{A}} \frac{\sqrt{2\log(2(L+2)/\delta_3)}}{\sqrt{m p_{A,Y}(a,y)} - \sqrt{2\log(2(L+2)/\delta_3)}} + \sum_{y \in \{0,1\}} \frac{\sqrt{2\log(2(L+2)/\delta_3)}}{\sqrt{m p_Y(y)} - \sqrt{2\log(2(L+2)/\delta_3)}}\right),$$

*and for an upper bound that covers both cases,*

$$\Delta = \frac{2(L+2)\log 2}{\epsilon} \frac{\sqrt{2\log(2(L+2)/\delta_3)}}{\sqrt{\eta m} - \sqrt{2\log(2(L+2)/\delta_3)}}.$$

We make the following remarks about the form of the bounds in Theorems 4 and 6.

1. The bounds are functions of two sample sizes: $n$, the number of data points that define the empirical dual (19), and $m$, the number of data points used to estimate $r(X)$ and $p_A(a)$ (in the MSP case) or $p_{A,Y}(a,y)$ (GEO case). The bounds have a familiar $1/\sqrt{n}$ dependence on $n$, and also on $m$ for the terms that correspond to estimation of $p_A(a)$ or $p_{A,Y}(a,y)$. The performance in estimating $r(X)$ is abstracted away by the error terms $E_{\mathrm{TV}}(m,\delta)$ and $E_{L_1}(m,\delta)$ defined in Assumptions 4 and 5.

2. The dimension $L$ of the dual variable $\lambda$, already no more than $2|\mathcal{A}|$ to begin with, enters mostly in logarithmic form.

3. The fairness tolerance $\epsilon$ and probability lower bound $\eta$ appear in the denominator (apart from the leading $\epsilon$ in Theorem 4). This agrees with the intuition that the problem becomes harder for stricter fairness constraints (smaller $\epsilon$) and smaller groups (smaller $\eta$). Some of the terms further specify the dependence on individual probabilities $p_A(a)$, $p_{A,Y}(a,y)$, $p_Y(y)$, which could be bounded by $\eta$ to simplify expressions.

### 5.3 Discussion and Basic Implications of Assumptions

We now elaborate upon the assumptions stated in Section 5.1.

#### 5.3.1 Assumption 2

Under this assumption, $\mu(X) = \lambda^T \mathbf{f}(X)$ is given by (16) (MSP) or (17) (GEO). In this case, $\mathbf{f}$ depends on $X$ only through $A$ and $r(X)$ and we will often use the notation $\mathbf{f}(A, r(X))$ to make this clear. Below we give expressions for $\mathbf{f}$ for future reference. For the MSP case (16), $\mathbf{f}$ has $|\mathcal{A}|$ components and the $a$th component is given by

$$f_a(X) = f_a(A) = \frac{\mathbf{1}(A = a)}{p_A(a)} - 1. \tag{24}$$

For the GEO case (17), $\mathbf{f}$ has $2|\mathcal{A}|$ components and the $(a, y)$ component is

$$f_{a,y}(X) = f_{a,y}(A, r(X)) = \begin{cases} \frac{1-r(X)}{p_Y(0)} \left( \frac{\mathbf{1}(A=a)}{p_{A \mid Y}(a \mid 0)} - 1 \right), & y = 0 \\ \frac{r(X)}{p_Y(1)} \left( \frac{\mathbf{1}(A=a)}{p_{A \mid Y}(a \mid 1)} - 1 \right), & y = 1. \end{cases} \tag{25}$$

For the estimate $\hat{\mathbf{f}}$ of $\mathbf{f}$, $p_A$ in (24) is replaced by its estimate $\hat{p}_A$, and $p_Y$, $p_{A \mid Y}$ (equivalently $p_{A,Y}$) in (25) are replaced by their estimates $\hat{p}_Y$, $\hat{p}_{A \mid Y}$ ($\hat{p}_{A,Y}$).

The proofs can be extended to the case in which $A$ is not known by also assuming a consistent estimator of the conditional probability $p_{A \mid X}$ in the MSP case or $p_{A \mid X,Y}$ in the GEO case and accounting for the error of this estimator.

#### 5.3.2 Assumption 3

This assumption is reasonable in that if a protected group is to be considered, it should represent a constant fraction of the population (and have non-negligible probabilities of being in classes 0 and 1). The boundedness of the estimated probabilities can be ensured by truncating them, i.e., setting $\hat{p}_A(a) \leftarrow \max\{\hat{p}_A(a), \eta\}$. If the minimum probability $p_A(a)$ or $p_{A,Y}(a, y)$ is known or imposed, $\eta$ can be set equal to this minimum probability. Note also that we must have $\eta \leq 1/|\mathcal{A}|$ for MSP and $\eta \leq 1/\big(2|\mathcal{A}|\big)$ for GEO, as otherwise $p_A$, $p_{A,Y}$ would sum to more than 1.

We further assume that the estimates $\hat{p}_A(a)$ and $\hat{p}_{A,Y}(a, y)$ are given by the corresponding empirical probabilities in a data set of size $m$. Each of these empirical probabilities is a binomial random variable with sample size parameter $m$ and scaled by $1/m$. Among many possible concentration inequalities, we make use of the following bound on the relative error. It follows from a Chernoff bound, as shown in Appendix A.2.1 for completeness.

**Lemma 7** *With probability at least $1 - \delta$, for any single $a \in \mathcal{A}$ or $(a, y) \in \mathcal{A} \times \{0, 1\}$,*

$$\left| \frac{p_A(a)}{\hat{p}_A(a)} - 1 \right| \leq \frac{\sqrt{2\log(2/\delta)}}{\sqrt{mp_A(a)} - \sqrt{2\log(2/\delta)}}, \qquad mp_A(a) > 2\log(2/\delta),$$

$$\left| \frac{p_{A,Y}(a, y)}{\hat{p}_{A,Y}(a, y)} - 1 \right| \leq \frac{\sqrt{2\log(2/\delta)}}{\sqrt{mp_{A,Y}(a, y)} - \sqrt{2\log(2/\delta)}}, \qquad mp_{A,Y}(a, y) > 2\log(2/\delta).$$

Note also that under Assumption 3, truncating the estimated probabilities at $\eta$ can only decrease the error and hence does not affect the bounds above.

#### 5.3.3 Assumption 4–6

In Assumptions 4 and 5, the properties of $E_{\mathrm{TV}}(m,\delta)$ and $E_{L_1}(m,\delta)$ imply that $D_{\mathrm{TV}}\big(\hat{r}(X) \mid A=a, r(X) \mid A=a\big)$ and $\mathbb{E}\left[|\hat{r}(X)-r(X)|\right]$ converge to zero in probability, similar to Assumption 6. This is true because for any deviation $E_{\mathrm{TV}}(m,\delta) > 0$ (similarly $E_{L_1}(m,\delta)$) and keeping $E_{\mathrm{TV}}(m,\delta)$ fixed, increasing $m$ requires increasing $1/\delta$ to compensate. Taking $m \to \infty$ thus drives $\delta$ (the probability of exceeding $E_{\mathrm{TV}}(m,\delta)$) to zero.

In Appendix A.2.2, it is shown that Assumption 6 implies the convergence in probability version of Assumption 5.

**Assumption 8** *The estimate $\hat{r}(X)$ converges to the population score $r(X)$ in $L_1$ norm: $\mathbb{E}\left[|\hat{r}(X)-r(X)|\right] \xrightarrow{p} 0$.*

**Lemma 8** *Assumption 6 implies Assumption 8.*

We list Assumption 8 separately as the proof of Lemma 17 below (for Theorem 5) requires only Assumption 8, not Assumption 6.

### 5.4 Proof Outlines

We begin with the proof of Theorem 6 as it contains elements that are reused in the proofs of Theorem 4 and 5.

#### 5.4.1 Asymptotic Dual Optimality (Theorem 6)

**Proof** We prove the theorem by deriving a uniform convergence bound on the absolute difference $|\hat{J}(\lambda) - J(\lambda)|$. Then if $\varepsilon$ is such a bound (that holds with high probability), we have

$$J(\hat{\lambda}) \le \hat{J}(\hat{\lambda}) + \varepsilon \le \hat{J}(\lambda^*) + \varepsilon \le J(\lambda^*) + 2\varepsilon, \tag{26}$$

where the second inequality is by definition of $\hat{\lambda}$.

Toward proving uniform convergence, we first establish that it suffices to solve the dual problem over a closed and bounded (and hence compact) feasible set. The same argument applies to both the population and empirical duals. Indeed, it always suffices to restrict to a sub-level set defined by the objective value of an initial solution. We take $\lambda = 0$ as the initial solution and consider $\{\lambda : J(\lambda) \le J(0)\}$ and $\{\lambda : \hat{J}(\lambda) \le \hat{J}(0)\}$. These sub-level sets are contained within an $\ell_1$ ball as proved in Appendix A.3.1.

**Lemma 9** *Given Assumption 7, define the $\ell_1$ ball*

$$\Lambda_0 = \left\{\lambda : \|\lambda\|_1 \le \frac{\log 2}{\epsilon}\right\}.$$

*Then we have $\{\lambda : J(\lambda) \le J(0)\} \subset \Lambda_0$ and $\{\lambda : \hat{J}(\lambda) \le \hat{J}(0)\} \subset \Lambda_0$.*

Henceforth we take $\Lambda_0$ to be the compact feasible set for the dual problem.

We then consider the supremum over $\Lambda_0$ of the absolute difference $|\hat{J}(\lambda) - J(\lambda)|$ as the quantity of interest for uniform convergence. We use the triangle inequality and separate suprema to decompose this into three terms:

$$\begin{aligned}\sup_{\lambda\in\Lambda_0}\left|\hat{J}(\lambda) - J(\lambda)\right| \leq & \sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n} g\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right| \\ & + \sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right] - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]\right| \\ & + \sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right] - \mathbb{E}\left[g\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\right]\right|. \end{aligned} \tag{27}$$

The first right-hand side quantity in (27) is the difference between the empirical average and expectation of the same quantity. The second difference is due to having $\hat{r}(X)$ instead of $r(X)$, and the third is due to having $\hat{\mathbf{f}}$ instead of $\mathbf{f}$.

The following three lemmas, proven in Appendix A.3, provide bounds on the three right-hand side terms in (27). Combining them with probabilities $\delta_1$, $\delta_2$, $\delta_3$ and including the factor of 2 from (26) completes the proof of the theorem.

**Lemma 10** *Under Assumptions 2, 3, 7 and with probability at least* $1-\delta$,

$$\begin{aligned}&\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n} g\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right| \\ &\qquad \leq \left(1+\frac{1}{\epsilon}\left(\frac{1}{\eta}-1\right)\right)(\log 2)\left(4\sqrt{\frac{2\log(2L)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}\right).\end{aligned}$$

**Lemma 11** *Under Assumptions 2, 3, 4, 7 and with probability* $1-\delta$,

$$\begin{aligned}&\sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right] - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]\right| \\ &\qquad \leq \left(1+\frac{1}{\epsilon}\left(\frac{1}{\eta}-1\right)\right)(\log 2)E_{\mathrm{TV}}(m,\delta).\end{aligned}$$

**Lemma 12** *Under Assumptions 2, 3, and 7 and with probability at least* $1-\delta$ *and* $m > (2/\eta)\log(2(L+2)/\delta)$, *in the MSP case,*

$$\begin{aligned}&\sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right] - \mathbb{E}\left[g\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\right]\right| \\ &\qquad \leq \frac{\log 2}{\epsilon}\sum_{a\in\mathcal{A}}\frac{\sqrt{2\log(2L/\delta)}}{\sqrt{mp_A(a)} - \sqrt{2\log(2L/\delta)}},\end{aligned}$$

*and in the GEO case,*

$$\sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]-\mathbb{E}\left[g\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\right]\right| \leq \frac{\log 2}{\epsilon}\left(\sum_{y\in\{0,1\}}\sum_{a\in\mathcal{A}}\frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_{A,Y}(a,y)}-\sqrt{2\log(2(L+2)/\delta)}} + \sum_{y\in\{0,1\}}\frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_Y(y)}-\sqrt{2\log(2(L+2)/\delta)}}\right),$$

*where* $L=\dim(\lambda)$.

■

### 5.4.2 Asymptotic Primal Feasibility (Theorem 4)

**Proof** By retracing the derivation of dual problems (14), (15) from the primal problem (7), it can be verified that the *empirical* primal corresponding to the empirical dual (19) is

$$\max_{r'} \quad -\frac{1}{n}\sum_{i=1}^{n} H_b\big(\hat{r}(x_i),r'(x_i)\big) \qquad \text{s. t.} \qquad \left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r'(x_i)\right|\leq\epsilon \quad \forall\, l, \tag{28}$$

where $l=a$ ranges over $\mathcal{A}$ in the MSP case and $l=(a,y)$ ranges over $\mathcal{A}\times\{0,1\}$ in the GEO case. Since $\hat{\lambda}$ optimizes the empirical dual, it follows from the discussion in Section 3 that the plug-in solution (22) satisfies the primal fairness constraints in (28). The task is to bound the amount by which the plug-in solution violates the population MSP (5) or GEO (6) constraints.

Using the definitions of $\mathbf{f}(A,r(X))$ for the MSP (24) and GEO (25) cases, it can be seen that constraints (5) and (6) are equivalent to

$$\left|\mathbb{E}\left[f_l(A,r(X))r'(X)\right]\right|\leq\epsilon \quad \forall\, l, \tag{29}$$

where $l$ ranges over the same values as in (28). Therefore by the triangle inequality, the violation of constraint $l$ in (29) is bounded by the difference

$$\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r'(x_i)-\mathbb{E}\left[f_l(A,r(X))r'(X)\right]\right|.$$

We apply the triangle inequality again to separate this difference into three terms that are analyzed below:

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r'(x_i) - \mathbb{E}\left[f_l(A,r(X))r'(X)\right]\right| \\
&\quad\le \left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r'(x_i) - \mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r'(X)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r'(X)\right] - \mathbb{E}\left[\hat{f}_l(A,r(X))r'(X)\right]\right| \\
&\qquad + \left|\mathbb{E}\left[\hat{f}_l(A,r(X))r'(X)\right] - \mathbb{E}\left[f_l(A,r(X))r'(X)\right]\right|. \qquad (30)
\end{aligned}$$

For the first right-hand side term in (30), we substitute in the plug-in solution (22) for $r'(X)$. To remove the dependence on $\hat{\lambda}$ (which is a function of the samples $i = 1,\dots,n$ to which it is fit), we consider a uniform bound over $\Lambda_0$:

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r^*\big(\hat{\lambda}^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r^*\big(\hat{\lambda}^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right| \\
&\le \sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r^*\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r^*\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right|.
\end{aligned}$$

The following bound is derived in Appendix A.4 using statistical learning theory tools similar to the proof of Lemma 10.

**Lemma 13** *Under Assumptions 2, 3, 7, with probability at least* $1-\delta$,

$$\begin{aligned}
&\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r^*\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r^*\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right| \\
&\quad\le \left(\frac{1}{\eta}-1\right)\left(\frac{4\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}} + \sqrt{\frac{2\log(2L/\delta)}{n}} + \frac{2}{\sqrt{n}}\right) \quad \forall\, l.
\end{aligned}$$

In the case of MSP, the second term in (30) is zero because $\mathbf{f}$ does not depend on its second argument $r(X)$. Using (24), the third term in (30) reduces as follows:

$$\begin{aligned}
\left|\mathbb{E}\left[\left(\frac{\mathbf{1}(A=a)}{\hat{p}_A(a)} - \frac{\mathbf{1}(A=a)}{p_A(a)}\right)r'(X)\right]\right| &= \left|\mathbb{E}\left[\left(\frac{p_A(a)}{\hat{p}_A(a)} - 1\right)r'(X)\mid A=a\right]\right| \\
&= \left|\frac{p_A(a)}{\hat{p}_A(a)} - 1\right|\left|\mathbb{E}\left[r'(X)\mid A=a\right]\right| \\
&\le \left|\frac{p_A(a)}{\hat{p}_A(a)} - 1\right| \\
&\le \frac{\sqrt{2\log(2L/\delta)}}{\sqrt{mp_A(a)} - \sqrt{2\log(2L/\delta)}}, \qquad (31)
\end{aligned}$$

where the first inequality is due to $|r'(X)| \leq 1$, and the second inequality from Lemma 7 holds with probability at least $1 - \delta/L$. By a union bound, (31) is true for all $a \in \mathcal{A}$ with probability at least $1 - \delta$ and $m$ large enough for the denominator to be positive.

In the GEO case, we prove in Appendix A.4 that the second and third terms in (30) are bounded as follows.

**Lemma 14** *In the GEO case, under Assumptions 2, 3, 5 and with probability at least* $1-\delta$*,*

$$\left|\mathbb{E}\left[\hat{f}_{a,y}(A, \hat{r}(X))r'(X)\right] - \mathbb{E}\left[\hat{f}_{a,y}(A, r(X))r'(X)\right]\right| \leq \left(\frac{1}{\eta} - 1\right) E_{L_1}(m, \delta) \quad \forall\, (a, y).$$

**Lemma 15** *In the GEO case, under Assumptions 2 and 3 and with probability at least* $1-\delta$ *and* $m > (2/\eta)\log(2(L+2)/\delta)$*,*

$$\begin{aligned}&\left|\mathbb{E}\left[\hat{f}_{a,y}(A, r(X))r'(X)\right] - \mathbb{E}\left[f_{a,y}(A, r(X))r'(X)\right]\right| \\ &\leq \frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_{A,Y}(a,y)} - \sqrt{2\log(2(L+2)/\delta)}} + \frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_Y(y)} - \sqrt{2\log(2(L+2)/\delta)}} \quad \forall\, (a, y).\end{aligned}$$

■

### 5.4.3 Asymptotic Primal Optimality (Theorem 5)

**Proof** We use the triangle inequality to bound the difference by the absolute sum of three differences:

$$\begin{aligned}&\left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); \hat{r}(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\lambda^{*T}\mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right| \\ &\quad\leq \left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); \hat{r}(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right)\right]\right| \\ &\qquad+ \left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right| \\ &\qquad+ \left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\mathbf{f}(A, r(X)); r(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\lambda^{*T}\mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right|.\end{aligned}$$

The first difference is due to having $\hat{r}(X)$ instead of $r(X)$ as the second argument to $r^*$, i.e., as the input score to the transformation. The second difference is due to having $\hat{\mathbf{f}}(A, \hat{r}(X))$ versus $\mathbf{f}(A, r(X))$, and the third to $\hat{\lambda}$ versus $\lambda^*$.

The following lemmas, proven in Appendix A.5, ensure that the three differences above converge to zero.

**Lemma 16** *Under Assumption 6,*

$$\left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); \hat{r}(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right)\right]\right| \xrightarrow{p} 0.$$

**Lemma 17** *Under Assumptions 2, 3, 7, and 8 (implied by Assumption 6),*

$$\left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T\mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right| \xrightarrow{p} 0.$$

**Lemma 18** *Under Assumptions 2, 3, 4, 7, 9,*

$$\left|\mathbb{E}\left[H_b\left(r(X), r^*\big(\hat{\lambda}^T\mathbf{f}(A, r(X)); r(X)\big)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\big(\lambda^{*T}\mathbf{f}(A, r(X)); r(X)\big)\right)\right]\right| \xrightarrow{p} 0.$$

■

The proof of Lemma 18 leverages the asymptotic dual optimality of $\hat{\lambda}$ as $n, m \to \infty$, implied by Theorem 6. In addition, we use the following assumption, where we define $s(\mu; r) = -\partial r^*(\mu; r)/\partial\mu$.

**Assumption 9** *For any empirical dual solution $\hat{\lambda}$ and any $\bar{\lambda}$ on the line segment between $\hat{\lambda}$ and a population dual solution $\lambda^*$ (i.e., $\bar{\lambda} = \alpha\hat{\lambda} + (1-\alpha)\lambda^*$ for $\alpha \in [0,1]$), there exists $\tau > 0$ such that*

$$\mathbb{E}\left[s\big(\bar{\lambda}^T\mathbf{f}(A, r(X)); r(X)\big)\left|\big(\hat{\lambda} - \lambda^*\big)^T\mathbf{f}(A, r(X))\right|^2\right] \geq \tau\mathbb{E}\left[\left|\big(\hat{\lambda} - \lambda^*\big)^T\mathbf{f}(A, r(X))\right|^2\right].$$

Assumption 9 is a form of strong convexity assumption on the first term $\mathbb{E}\left[g\big(\lambda^T\mathbf{f}(A, r(X)); r(X)\big)\right]$ in the population dual objective function, as will be seen in the proof of Lemma 18. The right-hand expectation in Assumption 9 is an $L_2$ norm between $\hat{\mu}(A, r(X)) = \hat{\lambda}^T\mathbf{f}(A, r(X))$ and $\mu^*(A, r(X)) = \lambda^{*T}\mathbf{f}(A, r(X))$, while the left-hand expectation is an $L_2$ norm weighted by $s(\mu; r)$. It can be seen from Figure 1a and verified using the expression in (91) that $s(\mu; r) \geq 0$ everywhere and $s(\mu; r) > 0$ for $r \in (0, 1)$. Hence, the assumption of a lower bound $\tau > 0$ is reasonable. However, whether Assumption 9 is satisfied depends on the distribution of the induced random variable $r(X)$ in a way that does not seem straightforward to characterize. Since $s(\mu; r)$ can be zero for $r = 0$ or $r = 1$, one requirement may be that $r(X)$ not have all of its probability mass at 0 and 1. It might also be possible in future work to prove Lemma 18 without Assumption 9.

Due in part to Assumption 9, in this work we do not pursue finite-sample guarantees or convergence rates to augment Theorem 5. Such bounds or rates would depend on the parameter $\tau$, which is not easy to interpret (and moreover may not be necessary). In addition, the proof of Lemma 16 is fairly involved and obtaining a rate for it does not appear straightforward.

## 6. Empirical Evaluation

This section discusses experimental evaluation of the proposed FST methods for MSP and GEO constraints and both the direct post-processing solution as well as the pre-processing extension.

### 6.1 Experimental Setup

We begin by describing the experimental setup, covering data sets, fairness methods, base classifiers, and metrics.

| | Adult | COMPAS | German | MEPS |
|---|---|---|---|---|
| number of instances | 45222 | 6167 | 1000 | 15830 |
| number of features | 13 | 10 | 20 | 41 |
| after one-hot encoding | 98 | 401 | 58 | 138 |
| percentage in positive class | 24.8 | 54.5 | 70.0 | 17.2 |
| protected attribute 1 | gender | gender | age | race |
| percentage in majority group | 67.5 | 81.0 | 85.1 | 64.3 |
| protected attribute 2 | race | race | | |
| percentage in majority group | 86.0 | 65.9 | | |

Table 2: Data set statistics

#### 6.1.1 Data Sets

Four data sets were used, the first three of which are standard in the fairness literature: 1) Adult Income, 2) ProPublica's COMPAS recidivism, 3) German credit risk, 4) Medical Expenditure Panel Survey (MEPS). Specifically, we used versions pre-processed by an open-source library for algorithmic fairness (Bellamy et al., 2018). Each data set was randomly split 10 times into training (75%) and test (25%) sets and all methods were subject to the same splits.

To facilitate comparison with other methods in Sections 6.2 and 6.3, we used binary-valued protected attributes and consider gender and race for both adult and COMPAS, age for German, and race for MEPS. The resulting data set statistics are shown in Table 2. In Section 6.4, we also evaluate FST on the Adult Income data set with both gender and race as protected attributes (i.e., four protected groups corresponding to the combinations).

#### 6.1.2 Methods Compared

Since FST is intended for post- and pre-processing, comparisons to other post- and pre-processing methods are most natural as they accommodate situations a)–c) in Section 1. For post-processing, we have chosen the method of Hardt et al. (2016) (HPS) and the reject option method of Kamiran et al. (2012), both as implemented by Bellamy et al. (2018), as well as the Wass-1 Post-Process $\hat{p}_S$ method (WPP) of Jiang et al. (2019). For pre-processing, the massaging and reweighing methods of Kamiran and Calders (2012) and the optimization method of Calmon et al. (2017) (OPP) were chosen. Among in-processing methods, meta-algorithms that work with essentially any base classifier can handle situation b). The reductions method of Agarwal et al. (2018) ('red') was selected from this class. We also compared to in-processing methods specific to certain types of classifiers, which do not allow for any of a)–c): fairness constraints (FC) (Zafar et al., 2017c), disparate mistreatment (DM) (Zafar et al., 2017a), and fair empirical risk minimization (FERM) (Donini et al., 2018). Lastly, availability of code was an important criterion.

The methods in the previous paragraph have various limitations, summarized by Table 3, that affect the design of the experiments. First, the post-processing methods (Hardt et al., 2016; Kamiran et al., 2012; Jiang et al., 2019, specifically the WPP variant for the last one) require knowledge of the protected attribute $A$ at test time. Accordingly, the experiments presented in Section 6.2 include $A$ in the features $X$ to make it available to all methods;

| method | pre | in | post | SP | EO | no $A$ at test time | scores | approx fair | any classifier |
|---|---|---|---|---|---|---|---|---|---|
| massage | ✓ | | | ✓ | | ✓ | ✓ | | ✓ |
| reweigh | ✓ | | | ✓ | | ✓ | ✓ | | ✓ |
| OPP | ✓ | | | ✓ | | ✓ | ✓ | ✓ | ✓ |
| HPS | | | ✓ | | ✓ | | | | ✓ |
| reject | | | ✓ | ✓ | $\star$ | | | ✓ | ✓ |
| WPP | | | ✓ | ✓ | | | ✓ | | ✓ |
| FC | | ✓ | | ✓ | | ✓ | ✓ | ✓ | |
| DM | | ✓ | | | ✓ | ✓ | ✓ | ✓ | |
| FERM | | ✓ | | | ✓ | ✓ | | ✓ | |
| reductions | | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| **FST** | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 3: Capabilities of methods in comparison. $\star$ refers to an extension implemented by Bellamy et al. (2018).

experiments without $A$ at test time (excluding Hardt et al., 2016; Kamiran et al., 2012; Jiang et al., 2019) are presented in Section 6.3. We also encountered computational problems with the methods of Calmon et al. (2017); Zafar et al. (2017a) and thus perform separate comparisons with FST on reduced feature sets, reported in Appendix C.

Three versions of FST were evaluated: direct post-processing (FSTpost), the pre-processing extension (FSTpre), and a second post-processing version (FSTbatch) that assumes that test instances can be processed in a batch rather than one by one. In this case, the fitting of the dual variables that parametrize FST (Section 4.2) can actually be done on test data since it does not depend on labels $y_i$ (and uses only predicted probabilities for $A$ if $A$ is unavailable at test time).

#### 6.1.3 Base Classifiers

We used $\ell_1$-regularized logistic regression (LR) and gradient boosted classification trees (GBM) from scikit-learn (Pedregosa et al., 2011) as base classifiers. These are used in different ways depending on the method: Post-processing methods operate on the scores produced by the base classifier, pre-processing methods train the base classifier after modifying the training data, and the reductions method repeatedly calls the base classification algorithm with different instance-specific costs. For FSTpre, the same base classifier is used both to obtain weights $w_i$ as well as to fit the re-weighted data. In Appendix C, we used linear SVMs (with the scaling of Platt, 1999, to output probabilities) to compare with FERM (Donini et al., 2018). We found it impractical to train nonlinear SVMs on the larger data sets for reductions and FERM since reductions needs to do so repeatedly and FERM uses a slower specialized algorithm. For a similar reason, 5-fold cross-validation to select parameters for LR (regularization parameter $C$ from $[10^{-4}, 10^4]$) and GBM (minimum number of samples per leaf from $\{5, 10, 15, 20, 30\}$) was done only once per training set. All other parameters were set to the scikit-learn defaults. The base classifier was then instantiated with the best parameter value for use by all methods.

#### 6.1.4 Metrics

Classification performance and fairness were evaluated using both score-based metrics (log loss, Brier score, and AUC for performance, differences in mean scores (MSP) and GEO for fairness) and binary label-based metrics (accuracy, differences in mean binary predictions (SP) and non-generalized EO). While FST optimizes log loss (recall from Section 2.1), we find that results for Brier score are highly similar and thus defer the log loss results to Appendix C.1. We account for the fact that the reductions method (Agarwal et al., 2018) returns a *randomized* classifier, i.e., a probability distribution over a set of classifiers. For the binary label-based metrics, we used the methods provided with the code[1] for reductions to compute the metrics. The score-based metrics were computed by evaluating the metric for each classifier in the distribution and then averaging, weighted by their probabilities.

### 6.2 Results with Exact Knowledge of Protected Attributes

Figures 2–6 show trade-offs between classification performance and fairness for the case where the features include the protected attribute $A$. We defer results on the German credit data set to Appendix C because its small size makes the results less conclusive. Appendix C also presents separate comparisons with FERM using linear SVMs, and with OPP and DM using reduced feature sets, as mentioned above.

Each of Figures 2–6 corresponds to one data set-protected attribute combination. The left two columns show score-based measures: Brier score in the leftmost column and AUC in the middle column versus MSP or GEO differences on the x-axis. The rightmost column shows binary label-based measures, namely accuracy vs. SP or EO differences. The rows correspond to combinations of base classifier (LR, GBM) and fairness measure targeted (SP, EO). Markers indicate mean values over the 10 splits, error bars indicate standard errors in the means, and Pareto-optimal points have been connected with line segments to ease visualization.

Considering first the score-based plots (left and middle columns), FSTpost and FSTbatch achieve trade-offs that are at least as good as all other methods, with a few slight exceptions involving GBMs (e.g., MEPS in Figure 6, AUC vs. MSP difference in Figure 2). In all cases, the advantage of FST lies in extending the Pareto frontiers farther to the left, attaining smaller MSP or GEO differences; this is especially apparent for GEO. FSTpre sometimes performs less well, e.g., with GBM on Adult (Figures 2 and 3) and MEPS (Figure 6). This is likely due to the additional step of approximating the transformed score $r'(x)$ with the output of a classifier fit to the pre-processed data, which incurs loss.

Turning to the binary label-based plots (right column), the trade-offs for FSTpost and FSTbatch generally coincide with or are close to the trade-offs of the best method, and are even sometimes the best, despite not optimizing for binary metrics beyond tuning the binarization threshold for accuracy. Again FSTpre with GBM is worse on Adult, but FSTpre with LR is a top performer on COMPAS (Figures 4 and 5). The main disadvantage of FST is that its trade-off curves may not extend as far to the left as other methods, in particular on Adult. This is the converse of its advantage for score-based metrics.

Among the existing methods, reductions is the strongest and also the most versatile, handling all cases that FST does. However, it is an in-processing method and far more

1. `https://github.com/microsoft/fairlearn`

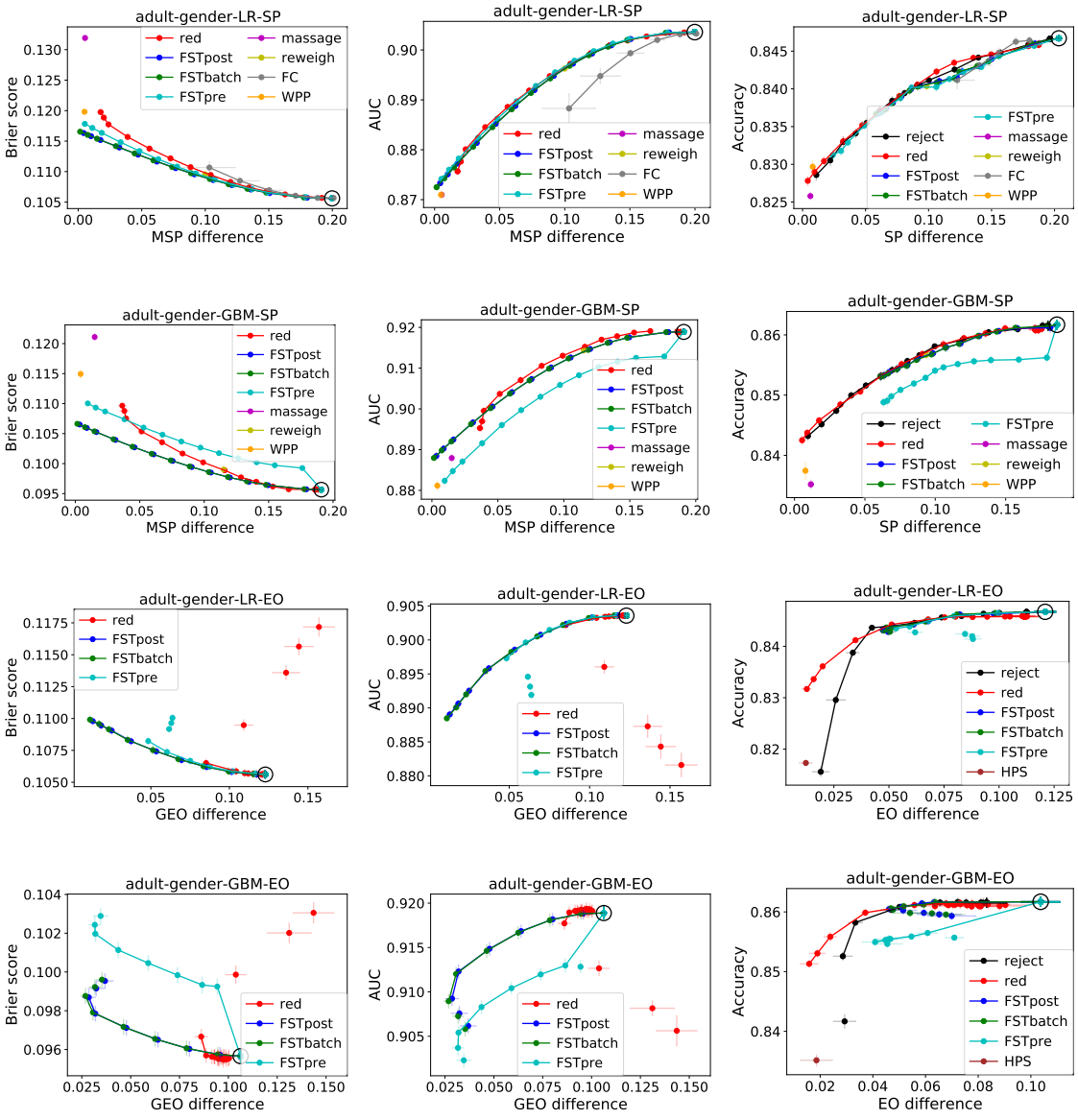



Figure 2: Trade-offs between fairness and classification performance on the Adult Income data set with gender as the protected attribute and the protected attribute included in the features. Pareto efficient points are connected by line segments to ease visualization. Horizontal and vertical bars represent standard errors in the means over 10 train-test splits. The point achieved by a classifier without fairness constraints is marked by a black circle.

computationally expensive, requiring an average of nearly 30 calls to the base classification algorithm compared to one for FSTpost, FSTbatch and two for FSTpre. Reductions also

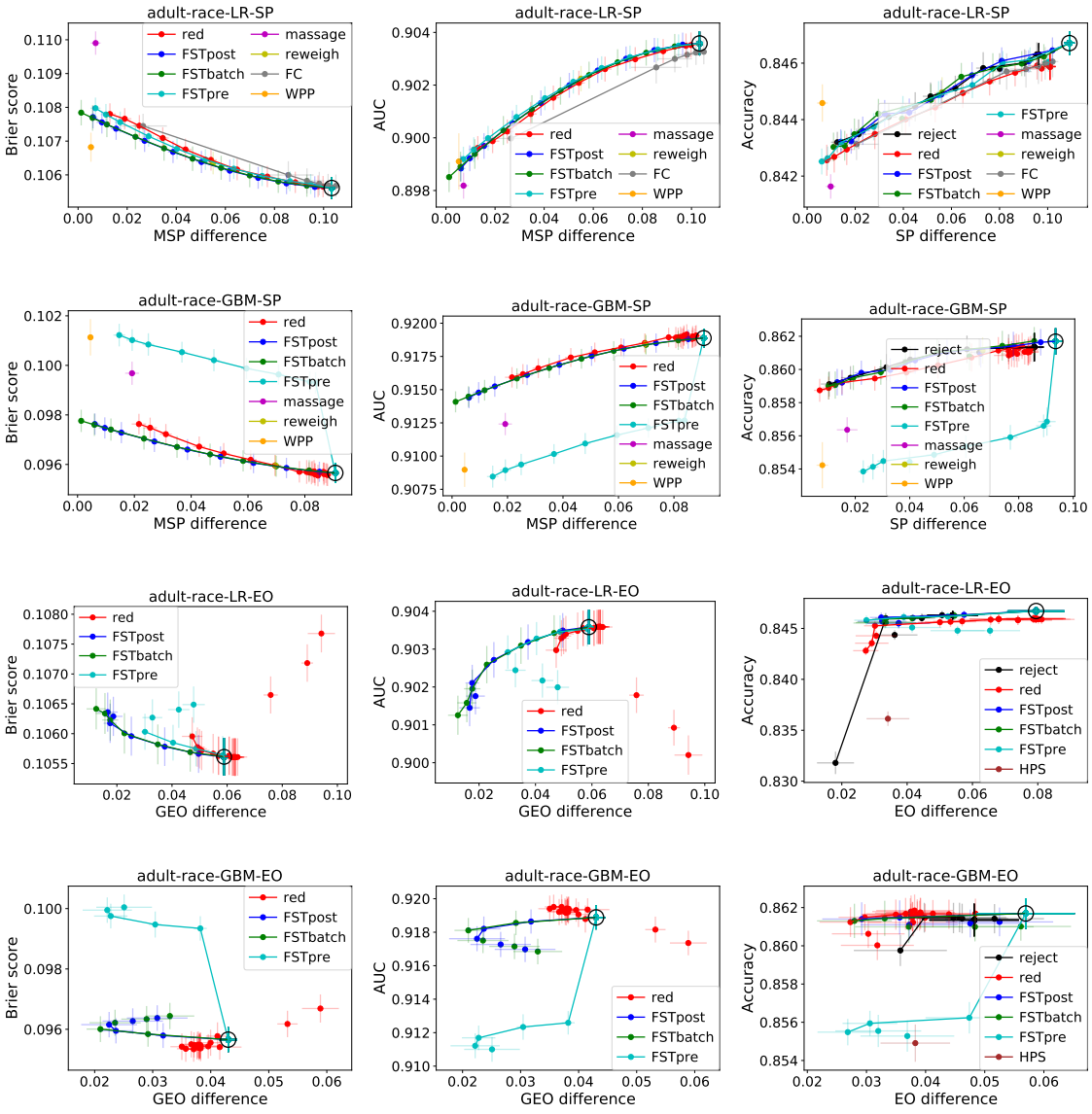



Figure 3: Trade-offs between fairness and classification performance on the Adult Income data set with race as the protected attribute and the protected attribute included in the features.

returns a randomized classifier, which may not be desirable in some applications. The other in-processing method shown in Figures 2–6 is FC, which applies only to the LR-SP rows (it is not compatible with GBM). It was not able to substantially reduce unfairness, particularly on COMPAS and MEPS and possibly due to the larger dimensionality of those data sets.

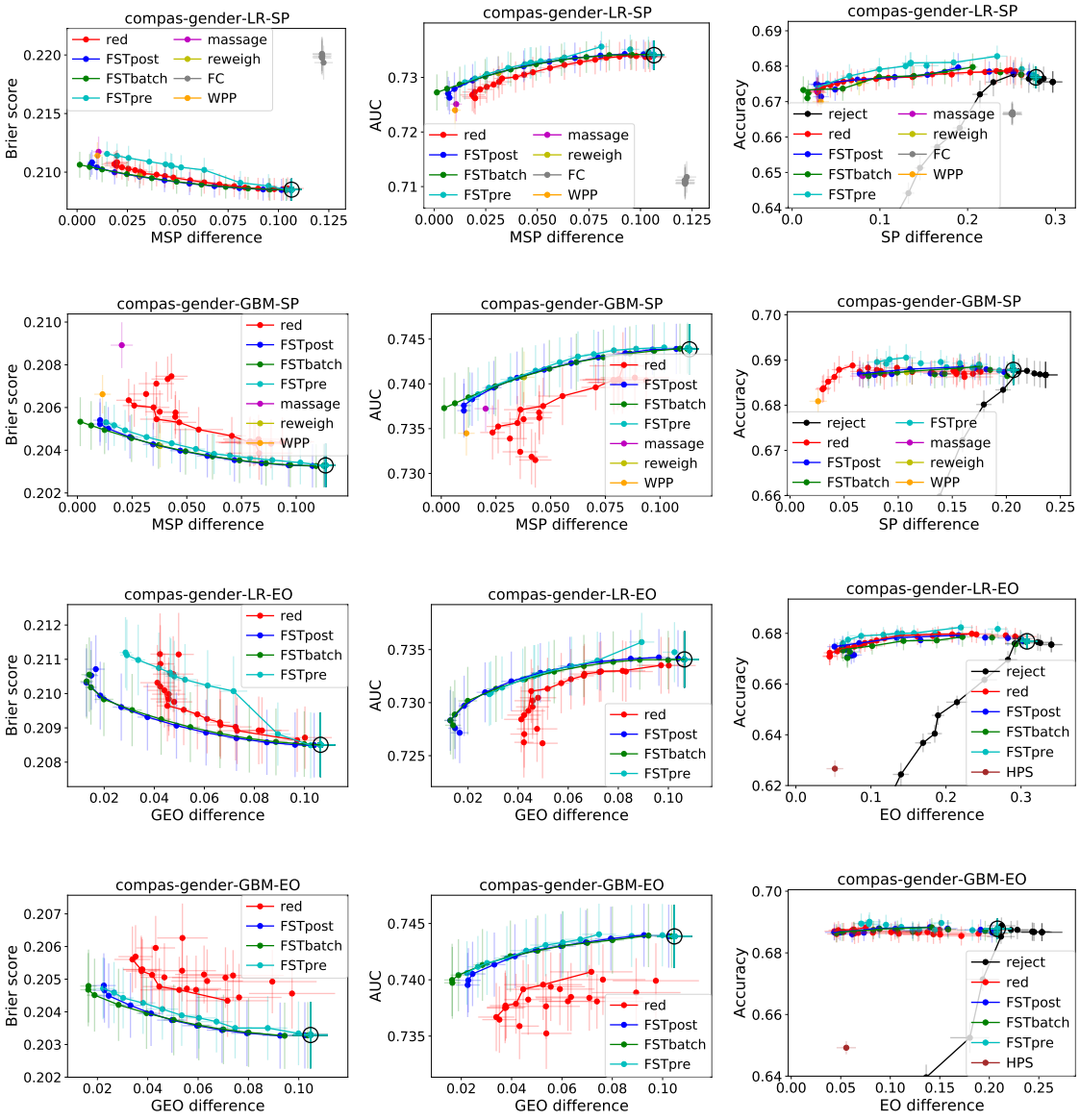



Figure 4: Trade-offs between fairness and classification performance on the COMPAS data set with gender as the protected attribute and the protected attribute included in the features.

The post-processing methods of Kamiran et al. (2012); Hardt et al. (2016) are not designed to output scores and hence are omitted from the score-based plots. Reject option (Kamiran et al., 2012) performs close to the best in many cases, but not on COMPAS-gender (Figure 4) and MEPS (Figure 6) and at small unfairness values. HPS is limited to EO, does not have a parameter to vary the trade-off, and is less competitive. WPP and the

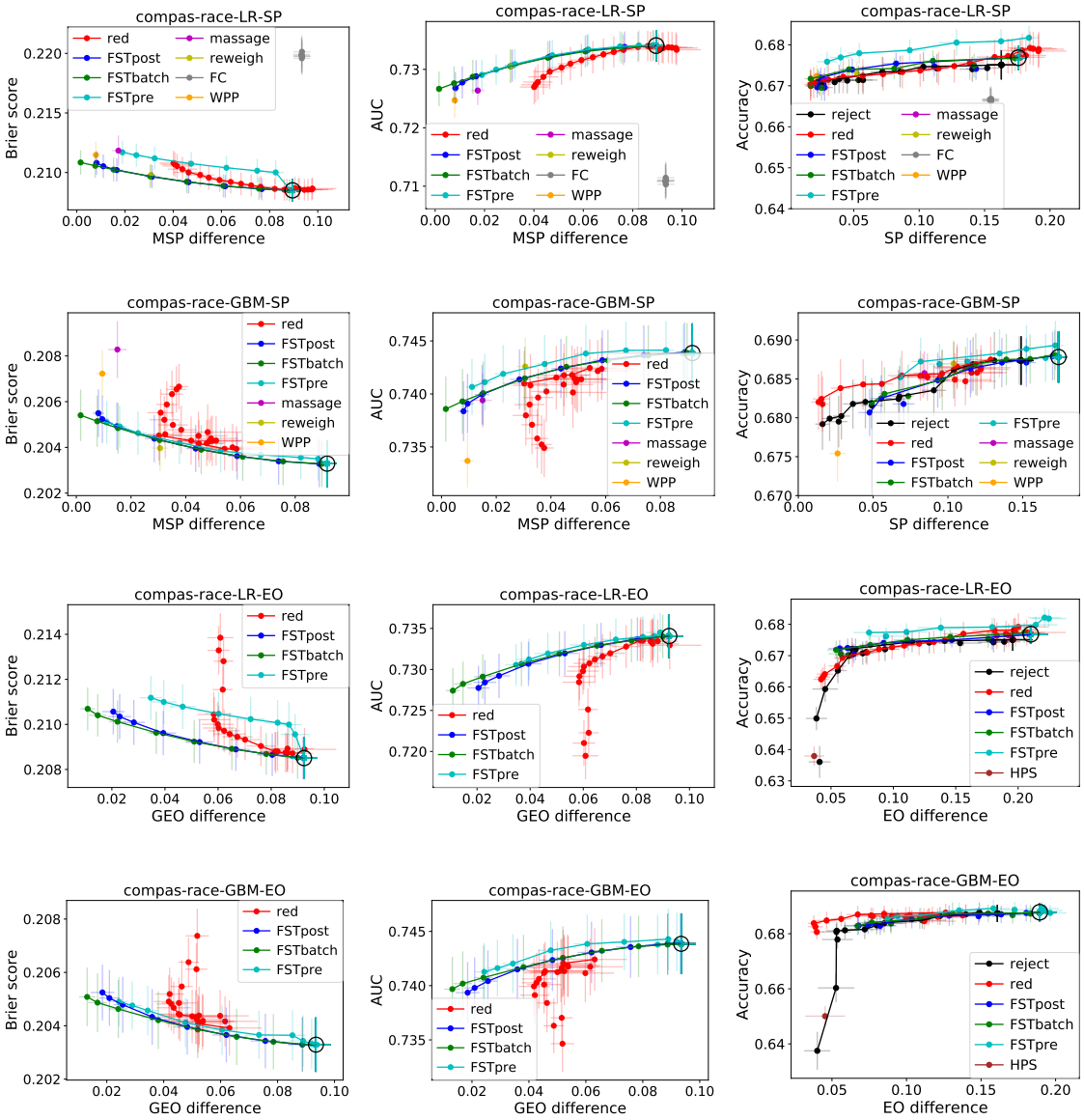



Figure 5: Trade-offs between fairness and classification performance on the COMPAS data set with race as the protected attribute and the protected attribute included in the features.

pre-processing methods of Kamiran and Calders (2012), massaging and reweighing, likewise do not have a trade-off parameter and are limited to SP. As also observed by Agarwal et al. (2018), massaging is often dominated by other methods while reweighing lies on the Pareto frontier but with substantial disparity. WPP results in low disparity but its classification performance (Brier score, AUC, or accuracy) is sometimes less competitive.

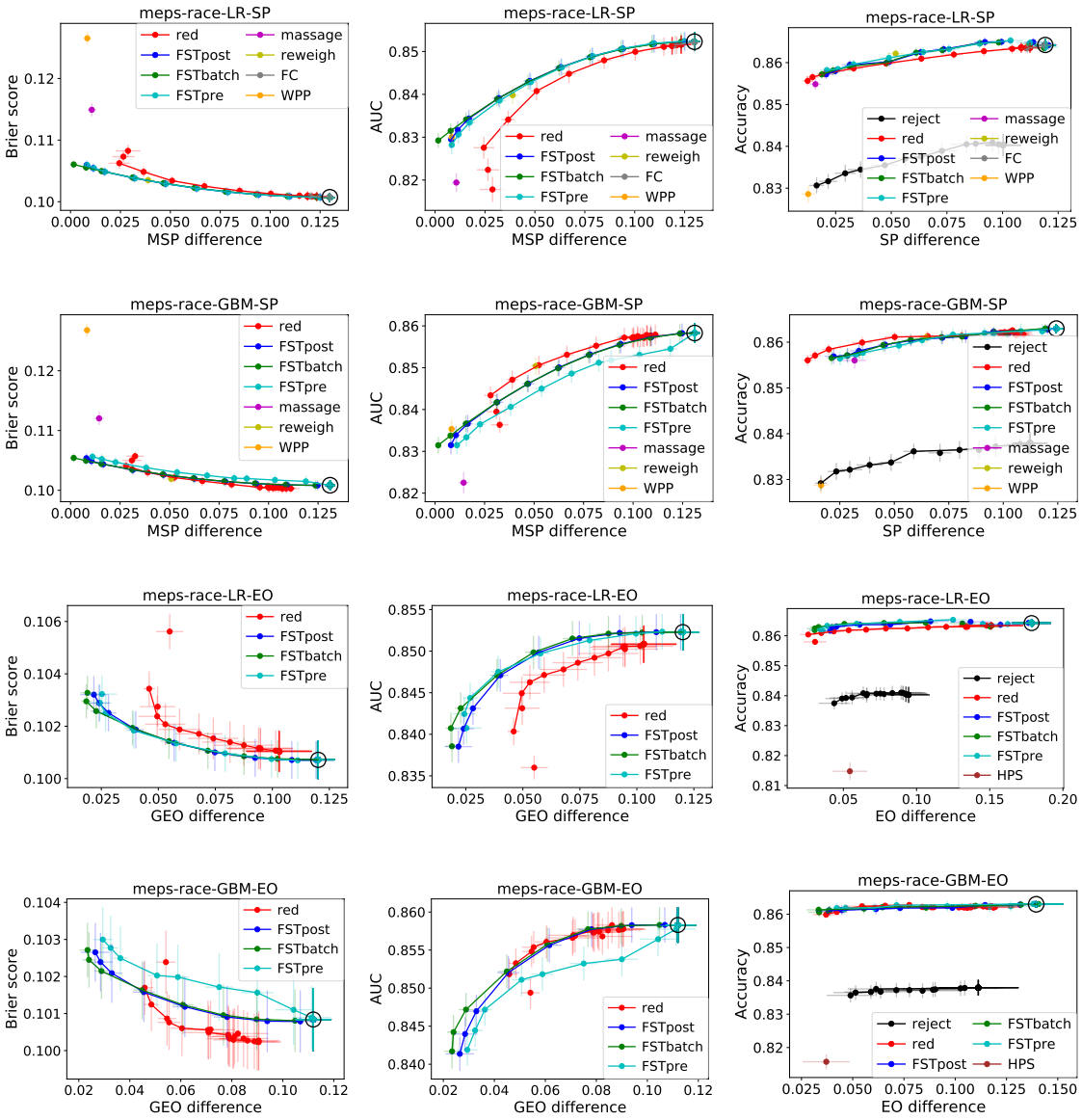



Figure 6: Trade-offs between fairness and classification performance on the MEPS data set with race as the protected attribute and the protected attribute included in the features.

### 6.3 Results with Inexact Knowledge of Protected Attributes

We now present results for the case where $A$ is excluded from the features and is not available at test time. We compare a smaller set of methods that can handle this case. For FST, we use the training data to train a probabilistic classifier for $A$ based on $X$ (for MSP) or $X, Y$ (for GEO), as discussed in Section 4.1. The same base classifier (LR or GBM) is used for

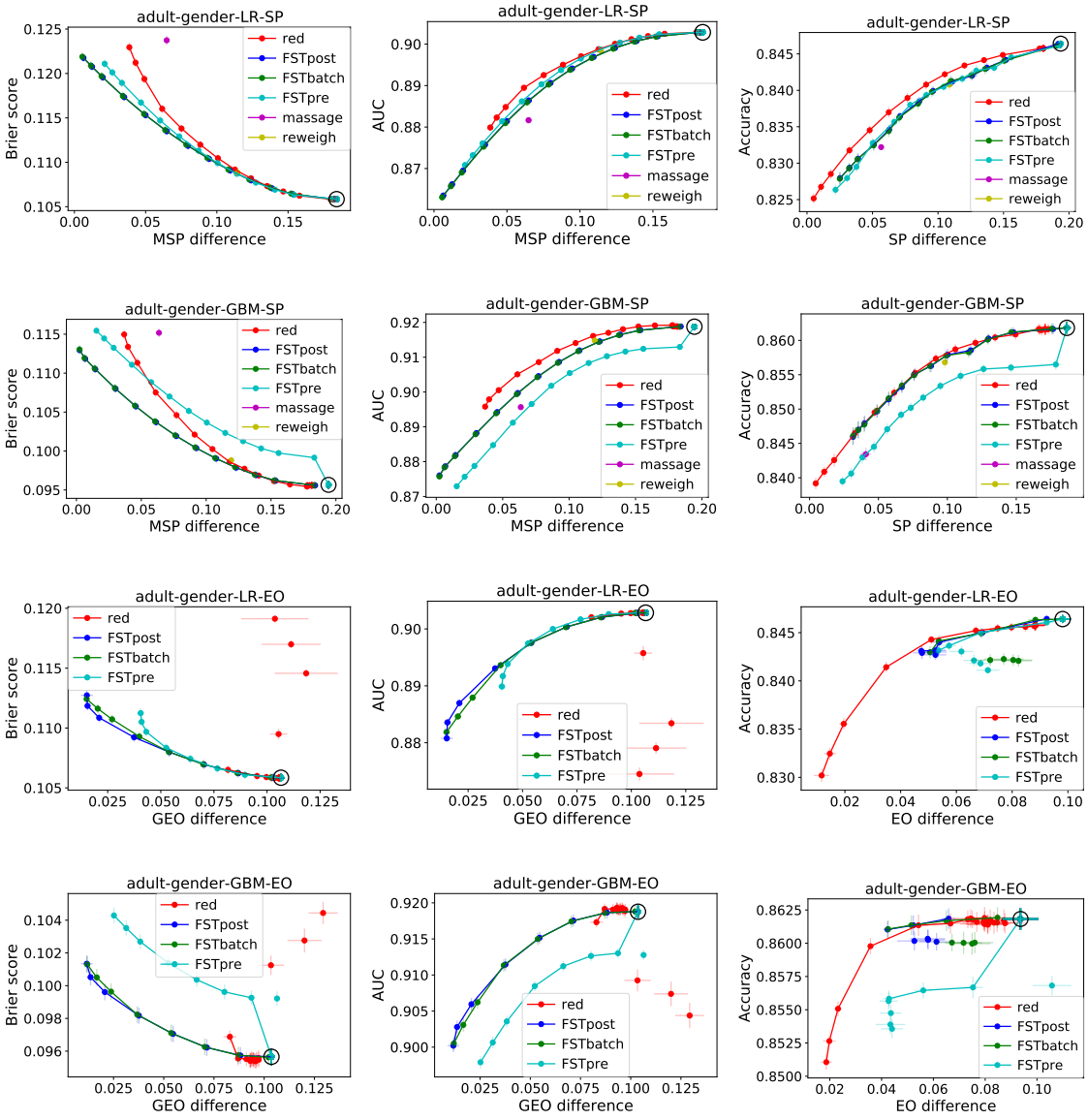



Figure 7: Trade-offs between fairness and classification performance on the Adult Income data set with gender as the protected attribute and the protected attribute excluded from the features.

this purpose. The classifier is used to approximate $p_{A \mid X}(a \mid x)$ in (14) or $p_{A \mid X,Y}(a \mid x, y)$ in (15), which are in turn used to compute $\mu(x)$ in both the fit and transform steps in Section 4.

The resulting trade-offs between classification performance and fairness are shown in Figures 7 and 8. Many of the patterns observed in Figures 2–6 reappear in Figure 7 (Adult-gender): FSTpost and FSTbatch dominate the Brier score column; FSTpre achieves worse

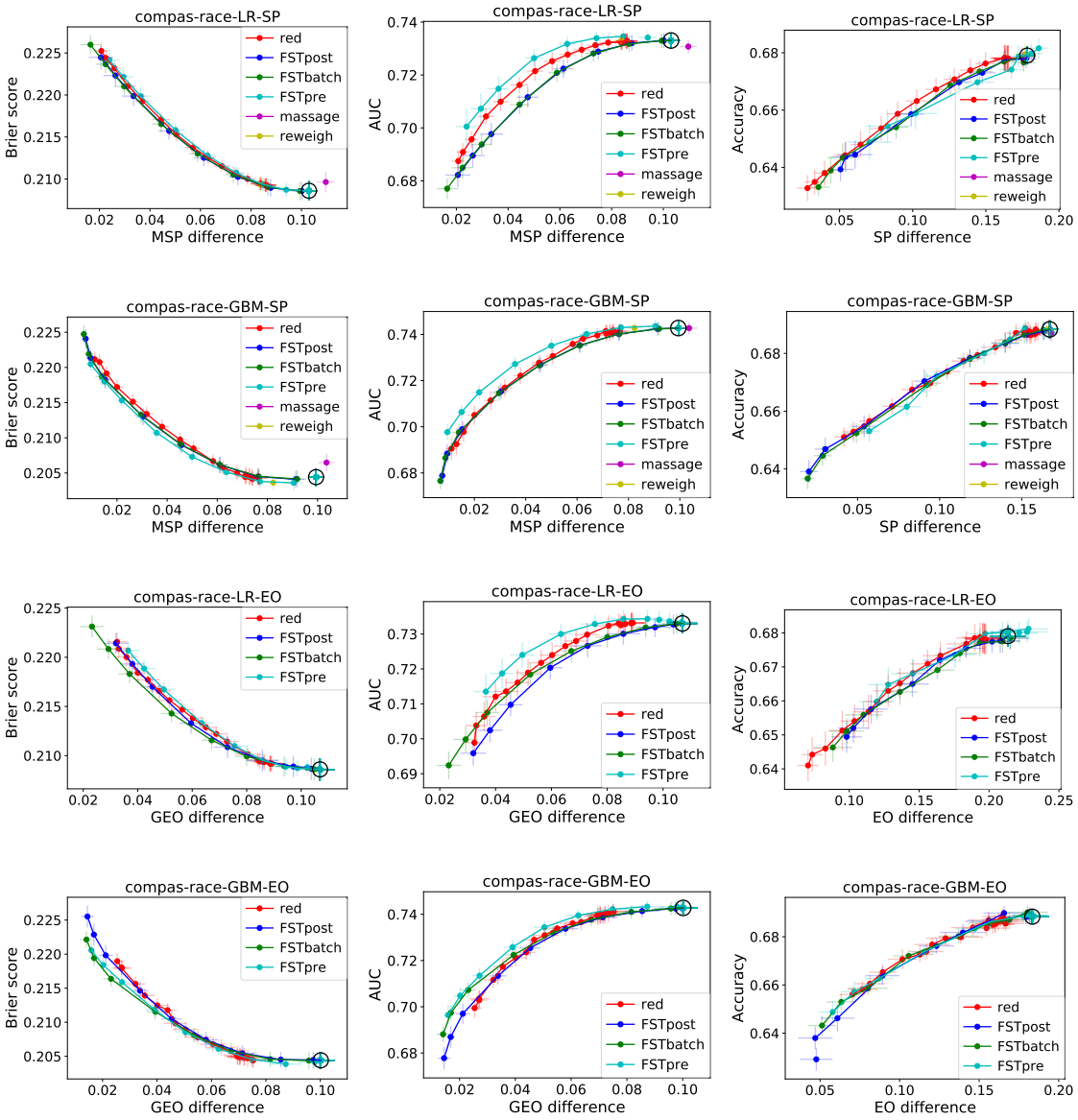



Figure 8: Trade-offs between fairness and classification performance on the COMPAS data set with race as the protected attribute and the protected attribute excluded from the features.

Brier scores, AUC, and accuracies with GBMs on Adult; FST achieves smaller score-based disparities while reductions achieves smaller binary prediction disparities (especially for GEO/EO); and reductions can obtain slightly better trade-offs with AUC and accuracy. All methods are more similar on COMPAS-race in Figure 8. In particular, FSTpre no longer lags and may have a slight advantage in the AUC column.

### 6.4 Results with More Than Two Protected Groups

The FST problem formulation also applies to non-binary protected groups. We evaluate this case using the Adult Income data set with both gender and race as protected attributes, giving rise to four protected groups (White males, Black males, White females, Black females). Here we do not compare FST with other methods as many of them do not handle more than two protected groups.

The results are shown in Figure 9 in the same style as Figures 2–8. With four protected groups, MSP and SP difference are computed as the largest difference in means between any two of the groups. Similarly, GEO and EO difference are computed as the largest (generalized) FPR or TPR difference between any two groups.

Figure 9 shows similar behavior to Figures 2 and 3 in particular. First, the pre-processing extension FSTpre results in worse Brier score, AUC, and accuracy values when applied to GBMs. Second, in the right-most column, the binary label-based measures of SP and EO difference are not reduced as much as the score-based MSP and GEO difference. In general, the SP and EO difference values are higher in Figure 9 than in Figures 2 and 3, due to having four groups (six possible pairs) instead of two. FSTpre does achieve significantly lower SP difference with LR than the post-processing versions (top right panel).

One difference compared to Figures 2 and 3 is that there is clearer separation between FSTpost, which is fit on training data, and FSTbatch, which is fit on test data. Specifically, for EO (bottom two rows), FSTbatch attains better trade-offs than FSTpost. A possible explanation is the greater difficulty of fairness generalization with effectively eight groups (four protected groups and two labels), which FSTbatch is able to sidestep to a degree.

## 7. Conclusion

This paper studied the problem of fair probabilistic classification, and specifically the transformation of predicted probabilities (scores) to satisfy fairness constraints with a linearity property (4) while minimizing cross-entropy (2) with respect to the input scores. We introduced a flexible solution method called FairScoreTransformer (FST), whose output can be used directly as post-processing and can also be adapted to pre-process training data. FairScoreTransformer takes advantage of a closed-form expression for the optimal transformed scores, a low-dimensional convex optimization for the Lagrange multiplier parameters, and an ADMM decomposition of this convex optimization to offer a computationally efficient solution. Theoretically, we showed in Section 5 that FST has asymptotic and finite-sample optimality and fairness consistency properties. Via a comprehensive set of experiments (Section 6 and Appendix C), we numerically demonstrated that FST is either as competitive or outperforms several existing fairness intervention mechanisms over a range of settings and data sets.

We note some limitations. First, FST inherently depends on well-calibrated classifiers that approximate $p_{Y \mid X}$ and, if necessary, $p_{A \mid X}$ or $p_{A \mid X,Y}$. This assumption of good calibration was made precise in Assumptions 4–6. A poorly calibrated model (e.g., due lack of samples) may lead to transformed scores that do not achieve the target fairness criteria. Second, thresholding the transformed scores may have an adverse impact on fairness guarantees, as seen in the right-hand columns throughout Figures 2–8. Third, the pre-processing extension of FST depends on the classifier trained on the original data and how

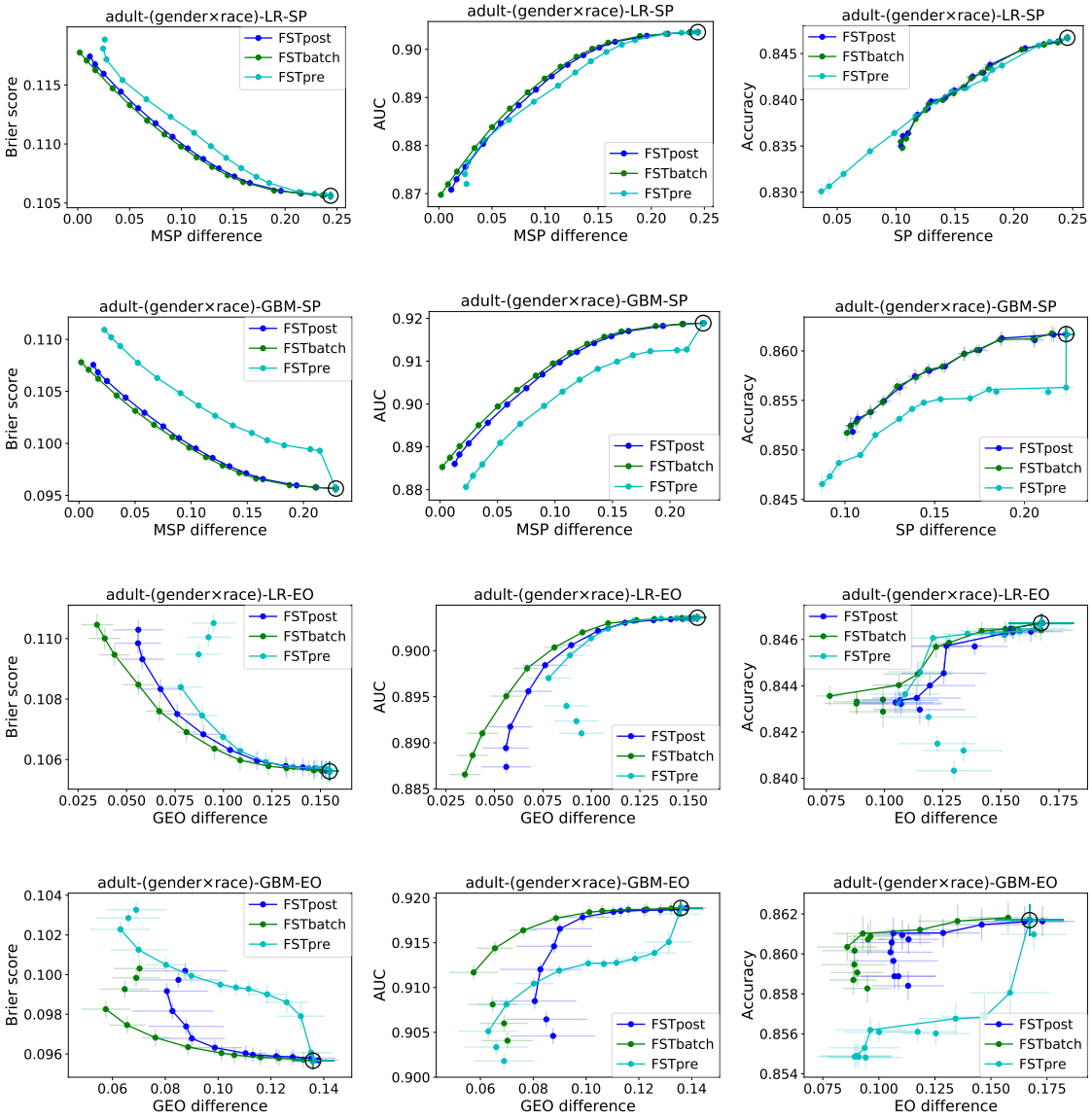



Figure 9: Trade-offs between fairness and classification performance on the Adult Income data set with both gender and race as protected attributes and the protected attributes included in the features.

well it approximates $p_{Y \mid X}$. The quality of this approximation limits subsequent classifiers trained on the re-weighted data. Finally, like most pre- and post-processing methods, the score transformation found by the FST is vulnerable to distribution shifts between training and deployment.

Future directions include: (1) characterizing the convergence rate of the ADMM iterations; (2) exploring alternative optimization algorithms for the empirical dual problem (19);

(3) adapting FairScoreTransformer to non-binary outcomes $Y$; (4) adapting FST to fairness criteria that are not based on conditional means of scores (e.g., calibration across groups as in Pleiss et al., 2017); (5) extending to other modalities such as text and images.

## Acknowledgments

F.P. Calmon would like to acknowledge support for this project from the National Science Foundation (NSF grant CIF-CAREER 1845852). All the authors thank the anonymous reviewers for their insightful comments during the review process, especially one reviewer whose attention to the proofs led to correction of flaws.

## Appendix A. Proofs

This appendix contains all proofs deferred from the main paper, organized by section and by theorem.

### A.1 Proofs for Section 3

Here we provide proofs of Propositions 1 and 3 and derivations for Table 1.

#### A.1.1 Proof of Proposition 1

**Proof** We manipulate the conditional mean scores as follows:

$$\begin{aligned}\mathbb{E}\big[r'(X) \mid \mathcal{E}_{lj}\big] &= \frac{\mathbb{E}\big[r'(X)\mathbf{1}((A,X,Y) \in \mathcal{E}_{lj})\big]}{\Pr(\mathcal{E}_{lj})} \\ &= \frac{\mathbb{E}\big[\mathbb{E}\big[r'(X)\mathbf{1}((A,X,Y) \in \mathcal{E}_{lj}) \mid X\big]\big]}{\Pr(\mathcal{E}_{lj})} \\ &= \frac{\mathbb{E}\big[r'(X)\Pr(\mathcal{E}_{lj} \mid X)\big]}{\Pr(\mathcal{E}_{lj})},\end{aligned}$$

where in the second line we have iterated expectations and then moved $r'(X)$ outside of the conditional expectation given $X$. Defining $\mu(X)$ according to (11), the Lagrangian (9) becomes

$$L(r', \lambda) = \mathbb{E}\big[\hat{r}(X)\log r'(X) + (1-\hat{r}(X))\log(1-r'(X)) - \mu(X)r'(X)\big] + \sum_{l=1}^{L} c_l\lambda_l. \quad (32)$$

It can be seen from (32) that the maximization with respect to the primal variable $r'(X)$ can be done independently for each $X = x$. Noting that $L(r', \lambda)$ is a concave function of $r'$ (sum of logarithmic and linear terms), a necessary and sufficient condition of optimality is that the partial derivatives with respect to each $r'(x)$ are equal to zero:

$$\frac{\hat{r}(x)}{r'(x)} - \frac{1-\hat{r}(x)}{1-r'(x)} - \mu(x) = 0 \quad \forall x \in \mathcal{X}. \quad (33)$$

This condition can be rearranged into the quadratic equation

$$\mu(x)r'(x)^2 - (1+\mu(x))r'(x) + \hat{r}(x) = 0,$$

whose solution is

$$r^*\big(\mu(x);\hat{r}(x)\big) = \begin{cases} \dfrac{1+\mu(x) - \sqrt{(1+\mu(x))^2 - 4\hat{r}(x)\mu(x)}}{2\mu(x)}, & \mu(x) \neq 0 \\ \hat{r}(x), & \mu(x) = 0, \end{cases}$$

after eliminating the root outside of the interval $[0, 1]$.

Lastly, it can be seen that the substitution of $r^*$ into the expectation in (32) yields $\mathbb{E}[g(\mu(X);\hat{r}(X))]$ where

$$g\big(\mu(x);\hat{r}(x)\big) \triangleq -H_b\Big(\hat{r}(x), r^*\left(\mu(x);\hat{r}(x)\right)\Big) - \mu(x)r^*\big(\mu(x);\hat{r}(x)\big).$$

■

### A.1.2 Proof of Proposition 3

**Proof** We first specify the exact correspondences between (5), (6) and (4). The MSP constraint (5) can be obtained from (4) by setting $J = 2$, $l = (a, \pm)$ for $a \in \mathcal{A}$ where $+$ corresponds to the $\leq \epsilon$ constraint and $-$ to the $\geq -\epsilon$ constraint, $L = 2|\mathcal{A}|$, $\mathcal{E}_{(a,\pm),1} = \{A = a\}$, $\mathcal{E}_{(a,\pm),2} = \Omega$ (the entire sample space), $c_l = \epsilon$, and $b_{(a,\pm),j} = \mp(-1)^j$. For the GEO constraint (6), set $J = 2$, $l = (a, y, \pm)$ for $a \in \mathcal{A}$, $y \in \{0, 1\}$ and the same $\pm$ correspondences, $L = 4|\mathcal{A}|$, $\mathcal{E}_{(a,y,\pm),1} = \{A = a, Y = y\}$, $\mathcal{E}_{(a,y,\pm),2} = \{Y = y\}$, $c_l = \epsilon$, and $b_{(a,y,\pm),j} = \mp(-1)^j$.

*Mean score parity constraints.* For MSP (5), let $\lambda_a^+$ and $\lambda_a^-$ respectively denote the Lagrange multipliers for the $\leq \epsilon$ and $\geq -\epsilon$ constraints for each $a \in \mathcal{A}$. With the correspondences identified above, the modifier $\mu(X, \lambda)$ becomes

$$\mu(X, \lambda) = \sum_{a \in \mathcal{A}} \left(\lambda_a^+ - \lambda_a^-\right) \left( \frac{p_{A \mid X}(a \mid X)}{p_A(a)} - \frac{\Pr(\Omega \mid X)}{\Pr(\Omega)} \right). \tag{34}$$

For $\epsilon > 0$, at most one of the constraints can be active for each $a$ in (5), and hence at optimality at most one of $\lambda_a^+$, $\lambda_a^-$ can be non-zero. We can therefore interpret $\lambda_a^+$, $\lambda_a^-$ as the positive and negative parts of a real-valued Lagrange multiplier $\lambda_a = \lambda_a^+ - \lambda_a^-$, as done in linear programming (Bertsimas and Tsitsiklis, 1997). Equation (34) can be rewritten as

$$\mu(X, \lambda) = \sum_{a \in \mathcal{A}} \lambda_a \frac{p_{A \mid X}(a \mid X)}{p_A(a)} - \sum_{a \in \mathcal{A}} \lambda_a. \tag{35}$$

If $A$ is included in the features $X$, then $p_{A \mid X}(a \mid X) = \mathbf{1}(a = A)$, where $A$ is the component of $X$ that is given, and (35) further simplifies to

$$\mu(X, \lambda) = \frac{\lambda_A}{p_A(A)} - \sum_{a \in \mathcal{A}} \lambda_a.$$

Interestingly, the only difference between the cases of including or excluding $A$ is that in the latter, (35) asks for $A$ to be inferred from the available features $X$, whereas in the former, $A$ can be used directly.

In the objective function of (13) we have

$$\sum_{l=1}^{L} c_l \lambda_l = \epsilon \sum_{a \in \mathcal{A}} \left(\lambda_a^+ + \lambda_a^-\right) = \epsilon \|\lambda\|_1 \tag{36}$$

upon recognizing that $(\lambda_a^+ + \lambda_a^-) = |\lambda_a|$. Combining this with (35), the dual problem for MSP is

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X); \hat{r}(X)\big)\right] + \epsilon \|\lambda\|_1 \\ \text{s. t.} \quad & \mu(X, \lambda) = \sum_{a \in \mathcal{A}} \lambda_a \frac{p_{A \mid X}(a \mid X)}{p_A(a)} - \sum_{a \in \mathcal{A}} \lambda_a. \end{aligned}$$

*Generalized equalized odds constraints.* For GEO (6), we similarly define Lagrange multipliers $\lambda^+_{a,y}$ and $\lambda^-_{a,y}$ for the $\leq \epsilon$ and $\geq -\epsilon$ constraints. The modifier $\mu(X)$ is given by

$$\begin{aligned}\mu(X,\lambda) &= \sum_{a\in\mathcal{A}}\sum_{y\in\{0,1\}}\left(\lambda^+_{a,y}-\lambda^-_{a,y}\right)\left(\frac{p_{A,Y\,|\,X}(a,y\,|\,X)}{p_{A,Y}(a,y)}-\frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)}\right)\\ &= \sum_{y\in\{0,1\}}\frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)}\sum_{a\in\mathcal{A}}\lambda_{a,y}\left(\frac{p_{A\,|\,X,Y}(a\,|\,X,y)}{p_{A\,|\,Y}(a\,|\,y)}-1\right), \end{aligned} \tag{37}$$

where we have similarly identified $\lambda_{a,y}=\lambda^+_{a,y}-\lambda^-_{a,y}$ and factored the joint distribution of $A,Y$. If $A$ is included in $X$, (37) simplifies to

$$\mu(X,\lambda)=\sum_{y\in\{0,1\}}\frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)}\left(\frac{\lambda_{A,y}}{p_{A\,|\,Y}(A\,|\,y)}-\sum_{a\in\mathcal{A}}\lambda_{a,y}\right).$$

Again, the difference between the two cases lies in whether $A$ must be inferred, this time from $X$ and $Y$. We also have an analogue to (36) where the summation and $\ell_1$ norm now run over all $(a,y)$. The dual problem for GEO is therefore

$$\begin{aligned}\min_{\lambda}\quad & \mathbb{E}\left[g\big(\mu(X,\lambda);\hat{r}(X)\big)\right]+\epsilon\|\lambda\|_1\\ \text{s.t.}\quad & \mu(X,\lambda)=\sum_{y\in\{0,1\}}\frac{p_{Y\,|\,X}(y\,|\,X)}{p_Y(y)}\sum_{a\in\mathcal{A}}\lambda_{a,y}\left(\frac{p_{A\,|\,X,Y}(a\,|\,X,y)}{p_{A\,|\,Y}(a\,|\,y)}-1\right).\end{aligned}$$

■

### A.1.3 Derivations for Table 1

As stated in Section 3.1, for the right-hand column of Table 1, we assume that the optimal transformed score $r^*(\mu(X);r(X))$ is thresholded at the cost-sensitive threshold $c$ to obtain a binary prediction, $\hat{Y}(X)=\mathbf{1}(r^*(\mu(X);r(X))>c)$. By virtue of the monotonicity of $r^*(\mu;r)$ in $r$ (Lemma 2 and Figure 1b), this is equivalent to thresholding $r(X)$ at a transformed threshold, which can be determined by setting $r^*=c$ and inverting (10) (see Appendix A.1.1 for the quadratic equation that leads to Equation 10). The result is

$$\hat{Y}(X)=\mathbf{1}\left(r(X)-c(1-c)\mu(X)>c\right), \tag{38}$$

i.e., an additive modification to the threshold that is proportional to $\mu(X)$.

We now discuss each row in Table 1 in turn. For the case of SP, Menon and Williamson (2018, Proposition 4) show that the classifier that minimizes (18) is given by

$$\hat{Y}^*(X)=\mathbf{1}\left(r(X)-\lambda(\bar{\eta}(X)-1/2)>c\right), \tag{39}$$

which is of the form $\mathbf{1}(h(X)>c)$ with $h(X)$ as given in the corresponding entry of Table 1. Two notes: (1) the $1/2$ in (39) comes from the equivalence of their MD criterion to a second cost-sensitive risk with weight $\bar{c}=1/2$ (Menon and Williamson, 2018, Lemma 2); (2) the

case where the threshold is met with equality is ignored for simplicity. On the other hand, for the thresholded optimal fair score (38) and the case of SP, the constraint in (14) gives

$$\mu(X) = \lambda_0 \left( \frac{1 - \bar{\eta}(X)}{p_A(0)} - 1 \right) + \lambda_1 \left( \frac{\bar{\eta}(X)}{p_A(1)} - 1 \right),$$

and hence

$$\hat{Y}(X) = \mathbf{1}\left( r(X) - c(1-c) \left( \lambda_0 \left( \frac{1 - \bar{\eta}(X)}{p_A(0)} - 1 \right) + \lambda_1 \left( \frac{\bar{\eta}(X)}{p_A(1)} - 1 \right) \right) > c \right). \tag{40}$$

This corresponds to the rightmost entry in the SP, $A$ not known row.

For the case of SP and $A$ known, (39) simplifies to (Menon and Williamson, 2018, Cor. 5)

$$\hat{Y}^*(X) = \begin{cases} \mathbf{1}\left( r(X) + \lambda/2 > c \right), & A = 0, \\ \mathbf{1}\left( r(X) - \lambda/2 > c \right), & A = 1, \end{cases}$$

while (40) becomes

$$\hat{Y}(X) = \begin{cases} \mathbf{1}\left( r(X) + c(1-c) p_A(1) \left( \frac{\lambda_1}{p_A(1)} - \frac{\lambda_0}{p_A(0)} \right) > c \right), & A = 0, \\ \mathbf{1}\left( r(X) - c(1-c) p_A(0) \left( \frac{\lambda_1}{p_A(1)} - \frac{\lambda_0}{p_A(0)} \right) > c \right), & A = 1. \end{cases}$$

The above two equations yield the SP, $A$ known row in Table 1.

For the case of EOpp, Menon and Williamson (2018, Proposition 6) specify the optimal classifier as follows:

$$\hat{Y}^*(X) = \mathbf{1}\left( \left( 1 - \frac{\lambda}{p_Y(1)} (\bar{\eta}(X) - 1/2) \right) r(X) > c \right), \tag{41}$$

where now $\bar{\eta}(X) = p_{A \mid X,Y}(1 \mid X, 1)$. For the thresholded transformed score in (38), an expression for $\mu(X)$ in the case of EOpp is needed. This is given by the constraint in (15) restricted to $y = 1$:

$$\mu(X) = \frac{r(X)}{p_Y(1)} \left( \lambda_0 \left( \frac{1 - \bar{\eta}(X)}{p_{A \mid Y}(0 \mid 1)} - 1 \right) + \lambda_1 \left( \frac{\bar{\eta}(X)}{p_{A \mid Y}(1 \mid 1)} - 1 \right) \right),$$

using the definitions of $r(X)$ and $\bar{\eta}(X)$ and dropping the second subscript $y = 1$ from $\lambda_{01}$, $\lambda_{11}$. Substituting into (38) yields

$$\hat{Y}(X) = \mathbf{1}\left( \left( 1 - \frac{c(1-c)}{p_Y(1)} \left( \lambda_0 \left( \frac{1 - \bar{\eta}(X)}{p_{A \mid Y}(0 \mid 1)} - 1 \right) + \lambda_1 \left( \frac{\bar{\eta}(X)}{p_{A \mid Y}(1 \mid 1)} - 1 \right) \right) \right) r(X) > c \right). \tag{42}$$

This establishes the third row in Table 1.

For the last case of EOpp and $A$ known, (41) and (42) simplify respectively to

$$\hat{Y}^*(X) = \begin{cases} \mathbf{1}\left( \left( 1 + \frac{\lambda}{2p_Y(1)} \right) r(X) > c \right), & A = 0, \\ \mathbf{1}\left( \left( 1 - \frac{\lambda}{2p_Y(1)} \right) r(X) > c \right), & A = 1, \end{cases}$$

$$\hat{Y}(X) = \begin{cases} \mathbf{1}\left( \left( 1 + \frac{c(1-c)}{p_Y(1)} p_{A \mid Y}(1 \mid 1) \left( \frac{\lambda_1}{p_{A \mid Y}(1 \mid 1)} - \frac{\lambda_0}{p_{A \mid Y}(0 \mid 1)} \right) \right) r(X) > c \right), & A = 0, \\ \mathbf{1}\left( \left( 1 - \frac{c(1-c)}{p_Y(1)} p_{A \mid Y}(0 \mid 1) \left( \frac{\lambda_1}{p_{A \mid Y}(1 \mid 1)} - \frac{\lambda_0}{p_{A \mid Y}(0 \mid 1)} \right) \right) r(X) > c \right), & A = 1. \end{cases}$$

### A.2 Proofs for Section 5.3

We prove two implications of the assumptions discussed in Section 5.3.

#### A.2.1 Proof of Lemma 7

**Proof** We prove the lemma for a generic probability $p$, which can be either $p_A(a)$ or $p_{A,Y}(a, y)$, and its empirical estimate $\hat{p}$. First we consider the event

$$\frac{p}{\hat{p}} - 1 > \varepsilon \quad \Longleftrightarrow \quad \hat{p} < \frac{p}{1+\varepsilon}$$

for $\varepsilon > 0$. A version of the Chernoff-Hoeffding theorem bounds the probability of this event as

$$\Pr\left(\hat{p} < \frac{p}{1+\varepsilon}\right) \leq \exp\left(-mD_{\mathrm{KL}}\left(\frac{p}{1+\varepsilon} \,\middle\|\, p\right)\right), \tag{43}$$

recalling the definition of Bernoulli KL divergence $D_{\mathrm{KL}}(p \,\|\, q)$ in (23). It can be shown by a second-order Taylor expansion that

$$D_{\mathrm{KL}}(p \,\|\, q) \geq \frac{(p-q)^2}{2\max\{p, q\}}.$$

Applying this to (43) yields

$$\Pr\left(\hat{p} < \frac{p}{1+\varepsilon}\right) \leq \exp\left(-\frac{mp^2(1/(1+\varepsilon)-1)^2}{2p}\right) = \exp\left(-\frac{mp}{2}\left(\frac{\varepsilon}{1+\varepsilon}\right)^2\right).$$

Setting the right-hand side equal to $\delta/2$ and solving for $\varepsilon$,

$$\begin{aligned} \frac{\varepsilon}{1+\varepsilon} &= \sqrt{\frac{2\log(2/\delta)}{mp}}, \\ \varepsilon &= \frac{\sqrt{2\log(2/\delta)}}{\sqrt{mp} - \sqrt{2\log(2/\delta)}}, \end{aligned} \tag{44}$$

which requires $mp > 2\log(2/\delta)$.

For the other direction

$$\frac{p}{\hat{p}} - 1 < -\varepsilon \quad \Longleftrightarrow \quad \hat{p} > \frac{p}{1-\varepsilon},$$

a similar calculation results in

$$\Pr\left(\hat{p} > \frac{p}{1-\varepsilon}\right) \leq \exp\left(-\frac{mp\varepsilon^2}{2(1-\varepsilon)}\right) < \exp\left(-\frac{mp}{2}\left(\frac{\varepsilon}{1+\varepsilon}\right)^2\right).$$

Hence by taking $\varepsilon$ as in (44), we have $|(p/\hat{p}) - 1| \leq \varepsilon$ with probability at least $1 - \delta$. ■

### A.2.2 Proof of Lemma 8

**Proof** Assumption 6 means that for every $\epsilon > 0$, we have

$$\Pr\left(\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)\right] \leq \epsilon\right) \to 1,$$

where the probability is with respect to the random estimator $\hat{r}$.

Suppose then that $\hat{r}$ satisfies $\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)\right] \leq 2\epsilon$. It can be shown via a second-order Taylor expansion that $D_{\mathrm{KL}}(r \,\|\, \hat{r}) \geq 2(r - \hat{r})^2$ for any $r, \hat{r} \in [0, 1]$. Hence we also have $\mathbb{E}\left[(r(X) - \hat{r}(X))^2\right] \leq \epsilon$. Furthermore, by the law of total expectations,

$$\begin{aligned}
&\mathbb{E}\left[|r(X) - \hat{r}(X)|\right] \\
&\quad = \Pr\left(|r(X) - \hat{r}(X)| \leq \sqrt{\epsilon}\right) \mathbb{E}\left[|r(X) - \hat{r}(X)| \,\middle|\, |r(X) - \hat{r}(X)| \leq \sqrt{\epsilon}\right] \\
&\qquad + \Pr\left(|r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right) \mathbb{E}\left[|r(X) - \hat{r}(X)| \,\middle|\, |r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right] \\
&\quad \leq \sqrt{\epsilon} + \Pr\left(|r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right) \mathbb{E}\left[|r(X) - \hat{r}(X)| \,\middle|\, |r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right] \\
&\quad \leq \sqrt{\epsilon} + \Pr\left(|r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right) \mathbb{E}\left[\frac{|r(X) - \hat{r}(X)|^2}{\sqrt{\epsilon}} \,\middle|\, |r(X) - \hat{r}(X)| > \sqrt{\epsilon}\right] \\
&\quad \leq \sqrt{\epsilon} + \frac{1}{\sqrt{\epsilon}} \mathbb{E}\left[|r(X) - \hat{r}(X)|^2\right] \\
&\quad \leq 2\sqrt{\epsilon},
\end{aligned}$$

where we have used the conditions $|r(X) - \hat{r}(X)| \leq \sqrt{\epsilon}$ and $|r(X) - \hat{r}(X)| > \sqrt{\epsilon}$ to obtain the first and second inequalities above, respectively. The third inequality follows from a similar total expectations decomposition of $\mathbb{E}\left[|r(X) - \hat{r}(X)|^2\right]$ and the last inequality from $\mathbb{E}\left[|r(X) - \hat{r}(X)|^2\right] \leq \epsilon$.

We have thus shown for any $\epsilon > 0$ that $\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)\right] \leq 2\epsilon$ implies $\mathbb{E}\left[|r(X) - \hat{r}(X)|\right] \leq 2\sqrt{\epsilon}$. Hence

$$\Pr\left(\mathbb{E}\left[|r(X) - \hat{r}(X)|\right] \leq 2\sqrt{\epsilon}\right) \geq \Pr\left(\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)\right] \leq 2\epsilon\right) \to 1,$$

as required for Assumption 5. ■

## A.3 Proofs for Asymptotic Dual Optimality

This section completes the proof of Theorem 6 (asymptotic dual optimality), as was outlined in Section 5.4.1.

### A.3.1 Proof of Lemma 9

**Proof** We prove the lemma only for the sub-level set of the population dual, $\{\lambda : J(\lambda) \leq J(0)\}$. The argument for the empirical dual $\hat{J}(\lambda)$ is entirely analogous. The inclusion in the $\ell_1$ ball $\Lambda_0$ is proven by showing that the first term $\mathbb{E}\left[g(\mu(X); r(X))\right]$ in $J(\lambda)$ is bounded from below by a constant. The expectation $\mathbb{E}\left[g(\mu(X); r(X))\right]$ is in fact the dual objective function corresponding to a primal problem in which $\epsilon = 0$, i.e., perfect fairness is required (zero MSP or GEO difference). By weak duality, $\mathbb{E}\left[g(\mu(X); r(X))\right]$ is lower bounded by the objective value of any primal solution satisfying perfect fairness. The set of constant

score functions $r'(X) = r'$ is a family of such solutions since their conditional means do not depend on $A$ or $Y$. The corresponding primal objective value is

$$-\mathbb{E}\left[H_b(r(X), r')\right] = \log r' \mathbb{E}\left[r(X)\right] + \log(1-r')\mathbb{E}\left[1 - r(X)\right] = m_Y \log r' + (1-m_Y)\log(1-r'),$$

where $m_Y = \mathbb{E}\left[Y\right] = \mathbb{E}\left[r(X)\right]$ since $r(X) = p_{Y \mid X}(1 \mid X)$. Maximizing this with respect to $r'$ yields

$$\mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] \geq \max_{r' \in [0,1]} -\mathbb{E}\left[H_b(r(X), r')\right] = -H_b(m_Y, m_Y) \geq -\log 2, \tag{45}$$

where the last inequality is due to binary entropy being bounded by $\log 2$.

Now for $\lambda$ such that $J(\lambda) \leq J(0)$, we have

$$\mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] + \epsilon\|\lambda\|_1 \leq \mathbb{E}\left[g\big(0; r(X)\big)\right] = \mathbb{E}\left[-H_b(r(X), r(X))\right],$$

using (12) and the fact that $r^*(0; r(x)) = r(x)$. Since binary entropy $H_b(r, r)$ is non-negative,

$$\mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] + \epsilon\|\lambda\|_1 \leq 0, \quad \lambda : J(\lambda) \leq J(0).$$

Combining this with (45) and dividing by $\epsilon$ (allowed by Assumption 7) gives the result. ■

### A.3.2 Auxiliary Lemmas

Here we establish bounds on functions that are used to prove subsequent lemmas.

**Lemma 19** *The function $g(\mu; r)$ is 1-Lipschitz in $\mu$ for any fixed $r \in [0, 1]$.*

**Proof** By the mean value theorem,

$$|g(\mu_2; r) - g(\mu_1; r)| = \left|\frac{\partial g(\mu; r)}{\partial \mu}\bigg|_{\bar{\mu}}\right| |\mu_2 - \mu_1|$$

for any $\mu_1 \leq \mu_2$ and some $\bar{\mu} \in [\mu_1, \mu_2]$. From (90), $|\partial g(\mu; r)/\partial \mu| = r^*(\mu; r) \leq 1$ and the result follows. ■

**Lemma 20** *Under Assumptions 2, 3, and 7,*

$$\left|g\big(\lambda^T \mathbf{f}(x); r(x)\big)\right| \leq \left(1 + \frac{1}{\epsilon}\left(\frac{1}{\eta} - 1\right)\right) \log 2 \triangleq \bar{G} \quad \forall\, \lambda \in \Lambda_0,\ x \in \mathcal{X}.$$

*The same bound holds if $\mathbf{f}$ is replaced by $\hat{\mathbf{f}}$ and $r(x)$ by $\hat{r}(x)$.*

**Proof** By the triangle inequality and Lemma 19,

$$\begin{aligned}\left|g\big(\lambda^T \mathbf{f}(x); r(x)\big)\right| &\leq \left|g\big(0; r(x)\big)\right| + \left|g\big(\lambda^T \mathbf{f}(x); r(x)\big) - g\big(0; r(x)\big)\right| \\ &\leq \left|g\big(0; r(x)\big)\right| + \left|\lambda^T \mathbf{f}(x)\right|.\end{aligned}$$

As in the proof of Lemma 9, $|g(0; r(x))| = H_b(r(x), r(x)) \leq \log 2$, while $|\lambda^T \mathbf{f}(x)| \leq \|\lambda\|_1 \|\mathbf{f}\|_\infty$ from Hölder's inequality. Thus

$$\left|g\left(\lambda^T \mathbf{f}(x); r(x)\right)\right| \leq \log 2 + \|\lambda\|_1 \|\mathbf{f}\|_\infty,$$

and combining this with Lemmas 9 and 21 yields the result. The same proof holds if $\mathbf{f}$ is replaced by $\hat{\mathbf{f}}$ and $r(x)$ by $\hat{r}(x)$ since Lemma 21 applies equally to $\hat{\mathbf{f}}$ and both $r(x), \hat{r}(x) \in [0, 1]$. ■

**Lemma 21** *Under Assumptions 2 and 3,*

$$\|\mathbf{f}(x)\|_\infty \leq \frac{1}{\eta} - 1, \quad \|\hat{\mathbf{f}}(x)\|_\infty \leq \frac{1}{\eta} - 1 \qquad \forall\, x \in \mathcal{X}.$$

**Proof** For the MSP case, expression (24) implies that

$$\|\mathbf{f}(X)\|_\infty \leq \max_{a \in \mathcal{A}} \max\left\{\frac{1}{p_A(a)} - 1, 1\right\} = \max\left\{\max_{a \in \mathcal{A}} \frac{1}{p_A(a)} - 1, 1\right\} \leq \frac{1}{\eta} - 1$$

by Assumption 3. The same is true for $\hat{\mathbf{f}}$.

For the GEO case, we infer from (25) that

$$\begin{aligned}
\|\mathbf{f}(X)\|_\infty &\leq \max_{a \in \mathcal{A},\, r \in [0,1]} \max\left\{\frac{1 - r(X)}{p_Y(0)}\left(\frac{1}{p_{A\,|\,Y}(a \mid 0)} - 1\right), \frac{r(X)}{p_Y(1)}\left(\frac{1}{p_{A\,|\,Y}(a \mid 1)} - 1\right), \right. \\
&\qquad \left. \frac{1 - r(X)}{p_Y(0)}, \frac{r(X)}{p_Y(1)}\right\} \\
&\leq \max_{a \in \mathcal{A}} \max\left\{\frac{1}{p_Y(0)}\left(\frac{1}{p_{A\,|\,Y}(a \mid 0)} - 1\right), \frac{1}{p_Y(1)}\left(\frac{1}{p_{A\,|\,Y}(a \mid 1)} - 1\right), \frac{1}{p_Y(0)}, \frac{1}{p_Y(1)}\right\} \\
&\leq \max\left\{\frac{1}{\eta} - \frac{1}{p_Y(0)}, \frac{1}{\eta} - \frac{1}{p_Y(1)}\right\} \\
&\leq \frac{1}{\eta} - 1,
\end{aligned}$$

where the second line results from four separate maximizations over $r$, the third line from applying Assumption 3 to the first two terms and using $|\mathcal{A}| \geq 2$ to drop the last two terms, and the last line from $p_Y(y) \leq 1$. Again the same is true for $\hat{\mathbf{f}}$. ■

### A.3.3 Proof of Lemma 10

**Proof** The quantity of interest is the supremum over a set $\Lambda_0$ of the difference between an empirical average and expectation of the same function $g\left(\lambda^T \hat{\mathbf{f}}(A, \hat{r}(X)); \hat{r}(X)\right)$, which is a function of parameters $\lambda \in \Lambda_0$ and random variables $A$, $\hat{r}(X)$. This difference is analogous to the difference between empirical and expected risks in statistical learning theory, with $g$ playing the role of the loss function and $\lambda$ the model parameters. The supremum can therefore be bounded using learning theory tools for establishing uniform convergence.

We consider in particular the Rademacher complexity of the function class $\{g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big) : \lambda \in \Lambda_0\}$:

$$R_n(\Lambda_0) = \mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g\big(\lambda^T\hat{\mathbf{f}}(A_i,\hat{r}(X_i));\hat{r}(X_i)\big)\right|\right], \tag{46}$$

where $\sigma_i$, $i = 1,\ldots,n$ are i.i.d. Rademacher random variables and the expectation is with respect to $\{\sigma_i, A_i, X_i\}_{i=1}^n$. Using the standard learning theory arguments of McDiarmid's inequality and symmetrization (see e.g., Liang, 2016; Duchi, and references therein), the supremum of the one-sided difference satisfies

$$\Pr\left(\sup_{\lambda\in\Lambda_0}\frac{1}{n}\sum_{i=1}^{n} g\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right] > 2R_n(\Lambda_0)+\varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2\bar{G}^2},\right), \tag{47}$$

where the only difference is that $g$ is bounded by $\bar{G}$ defined in Lemma 20 instead of the unit interval $[0,1]$, and hence $\bar{G}^2$ appears in the exponent above. A similar bound holds for the difference in the other direction. Setting the right-hand side of (47) equal to $\delta/2$ and applying the union bound, we therefore have

$$\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n} g\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big) - \mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right| \leq 2R_n(\Lambda_0)+\bar{G}\sqrt{\frac{2\log(2/\delta)}{n}} \tag{48}$$

with probability at least $1-\delta$.

The proof is completed by obtaining an upper bound on the Rademacher complexity $R_n(\Lambda_0)$. For this, we consider the *empirical* Rademacher complexity obtained by conditioning on $A_i = a_i$, $\hat{r}(X_i) = \hat{r}_i$ in (46):

$$\hat{R}_n(\Lambda_0) = \mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big)\right|\right],$$

where $\hat{\mathbf{f}}_i = \hat{\mathbf{f}}(a_i,\hat{r}_i)$ is also fixed. The Rademacher complexity $R_n(\Lambda_0)$ is then the expectation of $\hat{R}_n(\Lambda_0)$ over $\{A_i,\hat{r}(X_i)\}$. First we subtract and add $g(0;\hat{r}_i)$,

$$\hat{R}_n(\Lambda_0) = \mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big) - g\big(0;\hat{r}_i\big) + g\big(0;\hat{r}_i\big)\right)\right|\right],$$

and treat $g(0;\hat{r}_i)$ as a $\lambda$-independent translation of the function class. Recalling from the proof of Lemma 9 that $|g(0;\hat{r}_i)| = H_b(\hat{r}_i,\hat{r}_i) \leq \log 2$ and using Bartlett and Mendelson (2002, Thm. 12.5),

$$\hat{R}_n(\Lambda_0) \leq \mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big) - g\big(0;\hat{r}_i\big)\right)\right|\right] + \frac{\log 2}{\sqrt{n}}. \tag{49}$$

The first term in (49) is the empirical Rademacher complexity of a composition of functions, $\hat{\mathbf{f}}_i \mapsto \lambda^T \hat{\mathbf{f}}_i$ and $\mu \mapsto g(\mu; \hat{r}_i) - g(0; \hat{r}_i)$. By Lemma 19, the second function satisfies the conditions of the Lipschitz composition property of Ledoux and Talagrand (1991) (see also Bartlett and Mendelson, 2002, Thm. 12.4), with Lipschitz constant 1. Applying their lemma results in

$$\begin{aligned}\mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big)-g\big(0;\hat{r}_i\big)\right)\right|\right] &\leq 2\mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\lambda^T\hat{\mathbf{f}}_i\right|\right]\\ &\leq 2\mathbb{E}\left[\sup_{\lambda\in\mathcal{B}_1((\log 2)/\epsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\lambda^T\hat{\mathbf{f}}_i\right|\right],\end{aligned}$$

where the second inequality is due to $\Lambda_0$ being contained in the $\ell_1$ ball of radius $(\log 2)/\epsilon$, $\mathcal{B}_1((\log 2)/\epsilon)$. Since $\lambda^T\hat{\mathbf{f}}_i$ is now linear in $\lambda$, we may use the standard steps of restricting to the vertices of $\mathcal{B}_1((\log 2)/\epsilon)$ and applying Massart's finite class lemma to obtain

$$\mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big)-g\big(0;\hat{r}_i\big)\right)\right|\right] \leq 2\sup_{\lambda\in\mathcal{B}_1((\log 2)/\epsilon)}\left|\lambda^T\hat{\mathbf{f}}_i\right|\sqrt{\frac{2\log(2L)}{n}},$$

with $L = \dim(\lambda)$. Applying Hölder's inequality $\left|\lambda^T\hat{\mathbf{f}}_i\right| \leq \|\lambda\|_1\left\|\hat{\mathbf{f}}_i\right\|_\infty$ and Lemma 21,

$$\mathbb{E}\left[\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\left(g\big(\lambda^T\hat{\mathbf{f}}_i;\hat{r}_i\big)-g\big(0;\hat{r}_i\big)\right)\right|\right] \leq \frac{2\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}}. \tag{50}$$

Substituting (50) into (49) and taking expectations,

$$R_n(\Lambda_0) \leq \frac{2\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}} + \frac{\log 2}{\sqrt{n}} < 2\bar{G}\sqrt{\frac{2\log(2L)}{n}}, \tag{51}$$

where the last inequality comes from $1 < 2\sqrt{2\log(2L)}$.

Combining (48) and (51) yields the result. ■

### A.3.4 Proof of Lemma 11

**Proof** We regard $\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]$ as the expectation of a function of $A$ and induced random variable $\hat{r}(X)$, and $\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]$ as the expectation of the same function of $A$ and induced random variable $r(X)$. Lemma 20 asserts that $g(\lambda^T\hat{\mathbf{f}}(a,r(x));r(x))$ is a bounded (and continuous) function for all $\lambda \in \Lambda_0$ and $x \in \mathcal{X}$. Therefore the convergence in distribution stated in Assumption 4 can be used to bound the difference in expectations.

More concretely, we have with probability at least $1-\delta$,

$$\begin{aligned}
&\sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]-\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]\right| \\
&=\sup_{\lambda\in\Lambda_0}\left|\sum_{a\in\mathcal{A}}p_A(a)\left(\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(a,\hat{r}(X));\hat{r}(X)\big)\mid A=a\right]-\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(a,r(X));r(X)\big)\mid A=a\right]\right)\right| \\
&\le\sup_{\lambda\in\Lambda_0}\sum_{a\in\mathcal{A}}p_A(a)\sup_{x\in\mathcal{X}}\left|g\big(\lambda^T\hat{\mathbf{f}}(a,r(x));r(x)\big)\right|D_{\mathrm{TV}}\big(\hat{r}(X)\mid A=a,r(X)\mid A=a\big) \\
&\le\left(1+\frac{1}{\epsilon}\left(\frac{1}{\eta}-1\right)\right)(\log 2)\sum_{a\in\mathcal{A}}p_A(a)D_{\mathrm{TV}}\big(\hat{r}(X)\mid A=a,r(X)\mid A=a\big) \\
&\le\left(1+\frac{1}{\epsilon}\left(\frac{1}{\eta}-1\right)\right)(\log 2)E_{\mathrm{TV}}(m,\delta).
\end{aligned}$$

The first inequality bounds the difference in expectations by the product of the total variation distance and the supremum of the function $g$. We then apply Lemma 20 to obtain the second inequality and invoke Assumption 4. ■

### A.3.5 Proof of Lemma 12

**Proof** Let us first consider a single point $x\in\mathcal{X}$. Using the Lipschitz property of $g$ in Lemma 19,

$$\left|g\big(\lambda^T\hat{\mathbf{f}}(a,r(x));r(x)\big)-g\big(\lambda^T\mathbf{f}(a,r(x));r(x)\big)\right|\le\left|\lambda^T\big(\hat{\mathbf{f}}(a,r(x))-\mathbf{f}(a,r(x))\big)\right|.$$

Hölder's inequality then gives

$$\left|g\big(\lambda^T\hat{\mathbf{f}}(a,r(x));r(x)\big)-g\big(\lambda^T\mathbf{f}(a,r(x));r(x)\big)\right|\le\|\lambda\|_1\left\|\hat{\mathbf{f}}(a,r(x))-\mathbf{f}(a,r(x))\right\|_\infty. \tag{52}$$

Now for the expectations over $\mathcal{X}$, it follows from the triangle inequality, (52), and Lemma 9 that

$$\begin{aligned}
&\sup_{\lambda\in\Lambda_0}\left|\mathbb{E}\left[g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)\right]-\mathbb{E}\left[g\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\right]\right| \\
&\quad\le\sup_{\lambda\in\Lambda_0}\mathbb{E}\left[\left|g\big(\lambda^T\hat{\mathbf{f}}(A,r(X));r(X)\big)-g\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\right|\right] \\
&\quad\le\sup_{\lambda\in\Lambda_0}\|\lambda\|_1\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right] \\
&\quad\le\frac{\log 2}{\epsilon}\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right].
\end{aligned}\tag{53}$$

Next we obtain bounds on the expected $\ell_\infty$ norm in (53), considering the MSP and GEO cases separately. In the former case, from (24) we obtain

$$\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty=\max_{a\in\mathcal{A}}\mathbf{1}(A=a)\left|\frac{1}{\hat{p}_A(a)}-\frac{1}{p_A(a)}\right|=\left|\frac{1}{\hat{p}_A(A)}-\frac{1}{p_A(A)}\right|.$$

Hence

$$\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right] = \sum_{a\in\mathcal{A}} p_A(a)\left|\frac{1}{\hat{p}_A(a)}-\frac{1}{p_A(a)}\right| = \sum_{a\in\mathcal{A}}\left|\frac{p_A(a)}{\hat{p}_A(a)}-1\right|.$$

Applying Lemma 7 to each term, together with a union bound over $a\in\mathcal{A}$, gives

$$\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right] \le \sum_{a\in\mathcal{A}} \frac{\sqrt{2\log(2L/\delta)}}{\sqrt{mp_A(a)}-\sqrt{2\log(2L/\delta)}} \tag{54}$$

with probability at least $1-\delta$, noting that $|\mathcal{A}|=L$ here.

For the GEO case, we use (25), the definition $r(X)=p_{Y\mid X}(1\mid X)$, and the triangle inequality to obtain

$$\begin{aligned}
&\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty \\
&= \max_{y\in\{0,1\}} p_{Y\mid X}(y\mid X)\max_{a\in\mathcal{A}}\left|\mathbf{1}(A=a)\left(\frac{1}{\hat{p}_{A,Y}(a,y)}-\frac{1}{p_{A,Y}(a,y)}\right)-\left(\frac{1}{\hat{p}_Y(y)}-\frac{1}{p_Y(y)}\right)\right| \\
&\le \max_{y\in\{0,1\}} p_{Y\mid X}(y\mid X)\left(\left|\frac{1}{\hat{p}_{A,Y}(A,y)}-\frac{1}{p_{A,Y}(A,y)}\right|+\left|\frac{1}{\hat{p}_Y(y)}-\frac{1}{p_Y(y)}\right|\right) \\
&\le \sum_{y\in\{0,1\}} p_{Y\mid X}(y\mid X)\left(\left|\frac{1}{\hat{p}_{A,Y}(A,y)}-\frac{1}{p_{A,Y}(A,y)}\right|+\left|\frac{1}{\hat{p}_Y(y)}-\frac{1}{p_Y(y)}\right|\right).
\end{aligned}$$

Taking expectations with respect to $X$ (which includes $A$),

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right] &\le \sum_{y\in\{0,1\}}\sum_{a\in\mathcal{A}} p_{A,Y}(a,y)\left|\frac{1}{\hat{p}_{A,Y}(a,y)}-\frac{1}{p_{A,Y}(a,y)}\right| \\
&\quad + \sum_{y\in\{0,1\}} p_Y(y)\left|\frac{1}{\hat{p}_Y(y)}-\frac{1}{p_Y(y)}\right| \\
&= \sum_{y\in\{0,1\}}\sum_{a\in\mathcal{A}}\left|\frac{p_{A,Y}(a,y)}{\hat{p}_{A,Y}(a,y)}-1\right| + \sum_{y\in\{0,1\}}\left|\frac{p_Y(y)}{\hat{p}_Y(y)}-1\right|.
\end{aligned}$$

The final right-hand side is a sum of $2(|\mathcal{A}|+1)=L+2$ terms of the form in Lemma 7. Applying this lemma and a union bound over the $L+2$ events, we have with probability at least $1-\delta$,

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A,r(X))-\mathbf{f}(A,r(X))\right\|_\infty\right] &\le \sum_{y\in\{0,1\}}\sum_{a\in\mathcal{A}} \frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_{A,Y}(a,y)}-\sqrt{2\log(2(L+2)/\delta)}} \\
&\quad + \sum_{y\in\{0,1\}} \frac{\sqrt{2\log(2(L+2)/\delta)}}{\sqrt{mp_Y(y)}-\sqrt{2\log(2(L+2)/\delta)}}.
\end{aligned} \tag{55}$$

The proof is completed by substituting (54) or (55) into (53). ■

### A.4 Proofs for Asymptotic Primal Feasibility

This section completes the proof of Theorem 4 (asymptotic primal feasibility), following the outline in Section 5.4.2.

#### A.4.1 Proof of Lemma 13

**Proof** We follow the same steps and use the same learning theory tools as in the proof of Lemma 10. The function class is now

$$\left\{\hat{f}_l(A,\hat{r}(X))r^*\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big) : \lambda\in\Lambda_0\right\}$$

and we consider its Rademacher complexity, again denoted by $R_n(\Lambda_0)$. Below we note the differences with respect to the proof of Lemma 10.

1. Using Lemma 21 and the fact that $|r^*(\cdot)|\leq 1$, the class of functions is uniformly bounded by $(1/\eta)-1$ in absolute value. Thus $(1/\eta)-1$ replaces $\bar{G}$ in the last term in (48).

2. Considering the empirical Rademacher complexity fixes the factor $\hat{f}_l(a_i,\hat{r}(x_i))$. We again use Lemma 21 to bound this factor by $(1/\eta)-1$. We then focus on determining the empirical Rademacher complexity of $\left\{r^*\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big) : \lambda\in\Lambda_0\right\}$, scaled by $(1/\eta)-1$.

3. The translating function that is added and subtracted is $r^*(0;\hat{r}_i)$ instead of $g(0;\hat{r}_i)$. Since $|r^*(0;\hat{r}_i)|=|\hat{r}_i|\leq 1$, the last $(\log 2)/\sqrt{n}$ term in (49) is replaced by $1/\sqrt{n}$ (which will be scaled by $(1/\eta)-1$).

4. From (91), it can be verified that $|\partial r^*(\mu;\hat{r})/\partial\mu|\leq 1$. Hence, like $g(\mu;\hat{r})$, $r^*(\mu;\hat{r})$ is 1-Lipschitz in $\mu$.

5. After using the Lipschitz composition property, the resulting class of $\ell_1$-bounded linear functions is the same as in the proof of Lemma 10.

In summary, we obtain

$$R_n(\Lambda_0)\leq\left(\frac{1}{\eta}-1\right)\left(\frac{2\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}}+\frac{1}{\sqrt{n}}\right)$$

and with probability at least $1-\delta$,

$$\sup_{\lambda\in\Lambda_0}\left|\frac{1}{n}\sum_{i=1}^{n}\hat{f}_l(a_i,\hat{r}(x_i))r^*\big(\lambda^T\hat{\mathbf{f}}(a_i,\hat{r}(x_i));\hat{r}(x_i)\big)-\mathbb{E}\left[\hat{f}_l(A,\hat{r}(X))r^*\big(\lambda^T\hat{\mathbf{f}}(A,\hat{r}(X));\hat{r}(X)\big)\right]\right|$$
$$\leq\left(\frac{1}{\eta}-1\right)\left(\frac{4\log 2}{\epsilon}\left(\frac{1}{\eta}-1\right)\sqrt{\frac{2\log(2L)}{n}}+\frac{2}{\sqrt{n}}+\sqrt{\frac{2\log(2/\delta)}{n}}\right).$$

The proof is completed by dividing the probability $\delta$ over $l=1,\ldots,L$ and applying the union bound. ■

### A.4.2 Proof of Lemma 14

**Proof** In the case of GEO, the index $l = (a, y)$ for $a \in \mathcal{A}$ and $y \in \{0, 1\}$. Using (25), the quantity of interest is given by

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{f}_{a,y}(A, \hat{r}(X))r'(X)\right] - \mathbb{E}\left[\hat{f}_{a,y}(A, r(X))r'(X)\right]\right| \\
&\qquad = \left|\mathbb{E}\left[(r(X) - \hat{r}(X))\left(\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,y)} - \frac{1}{\hat{p}_Y(y)}\right)r'(X)\right]\right| \\
&\qquad \leq \mathbb{E}\left[\left|r(X) - \hat{r}(X)\right|\left|\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,y)} - \frac{1}{\hat{p}_Y(y)}\right|r'(X)\right].
\end{aligned}$$

Similar to the proof of Lemma 21, we have

$$\left|\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,y)} - \frac{1}{\hat{p}_Y(y)}\right| \leq \max_{a\in\mathcal{A}, y\in\{0,1\}} \max\left\{\frac{1}{\hat{p}_{A,Y}(a,y)} - \frac{1}{\hat{p}_Y(y)}, \frac{1}{\hat{p}_Y(y)}\right\} \leq \frac{1}{\eta} - 1,$$

and $|r'(X)| \leq 1$. Hence we obtain the further bounds

$$\begin{aligned}
\left|\mathbb{E}\left[\hat{f}_{a,y}(A, \hat{r}(X))r'(X)\right] - \mathbb{E}\left[\hat{f}_{a,y}(A, r(X))r'(X)\right]\right| &\leq \left(\frac{1}{\eta} - 1\right)\mathbb{E}\left[\left|r(X) - \hat{r}(X)\right|\right] \\
&\leq \left(\frac{1}{\eta} - 1\right)E_{L_1}(m, \delta),
\end{aligned}$$

where the last bound holds with probability $1 - \delta$ by Assumption 5. ■

### A.4.3 Proof of Lemma 15

**Proof** As in Lemma 14, the index $l = (a, y)$ for $a \in \mathcal{A}$ and $y \in \{0, 1\}$. We prove the lemma for $y = 0$; the case $y = 1$ is similar.

Using (25), the quantity of interest is given by

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{f}_{a,0}(A, r(X))r'(X)\right] - \mathbb{E}\left[f_{a,0}(A, r(X))r'(X)\right]\right| \\
&\qquad = \left|\mathbb{E}\left[(1 - r(X))\left(\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,0)} - \frac{\mathbf{1}(A=a)}{p_{A,Y}(a,0)} - \frac{1}{\hat{p}_Y(0)} + \frac{1}{p_Y(0)}\right)r'(X)\right]\right|.
\end{aligned}$$

Recalling that $1 - r(x) = p_{Y\,|\,X}(0 \mid x)$ and using Bayes' rule $p_{Y\,|\,X}(0 \mid x)p_X(x) = p_Y(0)p_{X\,|\,Y}(x \mid 0)$, this can be rewritten as

$$\begin{aligned}
&\left|\mathbb{E}\left[\hat{f}_{a,0}(A, r(X))r'(X)\right] - \mathbb{E}\left[f_{a,0}(A, r(X))r'(X)\right]\right| && (56) \\
&= \left|\mathbb{E}\left[p_Y(0)\left(\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,0)} - \frac{\mathbf{1}(A=a)}{p_{A,Y}(a,0)} - \frac{1}{\hat{p}_Y(0)} + \frac{1}{p_Y(0)}\right)r'(X)\,\middle|\, Y=0\right]\right| \\
&\leq \left|\mathbb{E}\left[p_Y(0)\left(\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,0)} - \frac{\mathbf{1}(A=a)}{p_{A,Y}(a,0)}\right)r'(X)\,\middle|\, Y=0\right]\right| + \left|\mathbb{E}\left[\left(1 - \frac{p_Y(0)}{\hat{p}_Y(0)}\right)r'(X)\,\middle|\, Y=0\right]\right|. && (57)
\end{aligned}$$

The second term in (57) is bounded similarly as in the proof of Theorem 4 for the MSP case:

$$\begin{aligned}\left|\mathbb{E}\left[\left(1-\frac{p_Y(0)}{\hat{p}_Y(0)}\right)r'(X)\;\middle|\;Y=0\right]\right| &= \left|1-\frac{p_Y(0)}{\hat{p}_Y(0)}\right|\left|\mathbb{E}\left[r'(X)\mid Y=0\right]\right| \\ &\leq \left|1-\frac{p_Y(0)}{\hat{p}_Y(0)}\right| \\ &\leq \frac{\sqrt{2\log(2/\delta)}}{\sqrt{mp_Y(0)}-\sqrt{2\log(2/\delta)}}, \end{aligned} \tag{58}$$

using $|r'(X)| \leq 1$ and Lemma 7 (with probability $1-\delta$). The first term in (57) is also similar:

$$\begin{aligned}&\left|\mathbb{E}\left[p_Y(0)\left(\frac{\mathbf{1}(A=a)}{\hat{p}_{A,Y}(a,0)}-\frac{\mathbf{1}(A=a)}{p_{A,Y}(a,0)}\right)r'(X)\;\middle|\;Y=0\right]\right| \\ &\quad= \left|\mathbb{E}\left[p_Y(0)p_{A\,|\,Y}(a\mid 0)\left(\frac{1}{\hat{p}_{A,Y}(a,0)}-\frac{1}{p_{A,Y}(a,0)}\right)r'(X)\;\middle|\;A=a,Y=0\right]\right| \\ &\quad= \left|\frac{p_{A,Y}(a,0)}{\hat{p}_{A,Y}(a,0)}-1\right|\left|\mathbb{E}\left[r'(X)\mid A=a,Y=0\right]\right| \\ &\quad\leq \left|\frac{p_{A,Y}(a,0)}{\hat{p}_{A,Y}(a,0)}-1\right| \\ &\quad\leq \frac{\sqrt{2\log(2/\delta)}}{\sqrt{mp_{A,Y}(a,0)}-\sqrt{2\log(2/\delta)}}, \end{aligned} \tag{59}$$

again using $|r'(X)| \leq 1$ and Lemma 7. Considering all $a \in \mathcal{A}$ and $y \in \{0,1\}$, there are $2|\mathcal{A}|+2 = L+2$ deviations to control in (58), (59), and hence we divide the probability $\delta$ by $L+2$ and apply the union bound. The result then follows from substituting (58), (59) into (57). ■

### A.5 Proofs for Asymptotic Primal Optimality

We now complete the proof of Theorem 5 as outlined in Section 5.4.3.

#### A.5.1 Proof of Lemma 16

**Proof** Let $\hat{\mu}(x) = \hat{\lambda}^T\hat{\mathbf{f}}(a,\hat{r}(x))$. Using the definition of binary cross-entropy in (2), the difference in question can be written as

$$\left|\mathbb{E}\left[r(X)\log\frac{r^*\big(\hat{\mu}(X);r(X)\big)}{r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}+(1-r(X))\log\frac{1-r^*\big(\hat{\mu}(X);r(X)\big)}{1-r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right]\right|. \tag{60}$$

For $\hat{\mu}(X)=0$, $r^*$ is an identity mapping and (60) reduces to

$$\left|\mathbb{E}\left[r(X)\log\frac{r(X)}{\hat{r}(X)}+(1-r(X))\log\frac{1-r(X)}{1-\hat{r}(X)}\right]\right| = \mathbb{E}\left[D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))\right].$$

Hence Assumption 6 is a necessary condition for (60) to converge to zero in probability. We will show that Assumption 6 is sufficient as well.

Toward this end, we apply the triangle inequality to bound (60) as follows:

$$\begin{aligned}
&\left|\mathbb{E}\left[r(X)\log\frac{r^*\big(\hat{\mu}(X);r(X)\big)}{r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}+(1-r(X))\log\frac{1-r^*\big(\hat{\mu}(X);r(X)\big)}{1-r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right]\right| \\
&\quad\le \mathbb{E}\left[\left|r(X)\log\frac{r^*\big(\hat{\mu}(X);r(X)\big)}{r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}+(1-r(X))\log\frac{1-r^*\big(\hat{\mu}(X);r(X)\big)}{1-r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right|\right] \\
&\quad\le \mathbb{E}\left[r(X)\left|\log\frac{r^*\big(\hat{\mu}(X);r(X)\big)}{r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right|+(1-r(X))\left|\log\frac{1-r^*\big(\hat{\mu}(X);r(X)\big)}{1-r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right|\right] \\
&\quad\le \mathbb{E}\left[r(X)\sup_{\mu}\left|\log\frac{r^*\big(\mu;r(X)\big)}{r^*\big(\mu;\hat{r}(X)\big)}\right|+(1-r(X))\sup_{\mu}\left|\log\frac{1-r^*\big(\mu;r(X)\big)}{1-r^*\big(\mu;\hat{r}(X)\big)}\right|\right].
\end{aligned}$$

The last inequality results from replacing $\hat{\mu}(X)$ with the supremum over $\mu$ for a given $r(x)$, $\hat{r}(x)$. Applying Lemma 22 and defining

$$\bar{D}(r\,\|\,\hat{r})=r\left|\log\frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}\right|+(1-r)\left|\log\frac{1-\sqrt{r}}{1-\sqrt{\hat{r}}}\right|$$

results in

$$\left|\mathbb{E}\left[r(X)\log\frac{r^*\big(\hat{\mu}(X);r(X)\big)}{r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}+(1-r(X))\log\frac{1-r^*\big(\hat{\mu}(X);r(X)\big)}{1-r^*\big(\hat{\mu}(X);\hat{r}(X)\big)}\right]\right|\le\mathbb{E}\left[\bar{D}(r(X)\,\|\,\hat{r}(X))\right]. \tag{61}$$

We have thus eliminated $\hat{\mu}(X)$ from the expectation.

We proceed to bound the right-hand side of (61). Using the law of total expectations, for any $\epsilon>0$,

$$\begin{aligned}
&\mathbb{E}\left[\bar{D}(r(X)\,\|\,\hat{r}(X))\right] \\
&\quad=\Pr\big(D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))\le\epsilon\big)\mathbb{E}\left[\bar{D}(r(X)\,\|\,\hat{r}(X))\,\middle|\,D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))\le\epsilon\right] \\
&\qquad+\Pr\big(D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))>\epsilon\big)\mathbb{E}\left[\bar{D}(r(X)\,\|\,\hat{r}(X))\,\middle|\,D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))>\epsilon\right] \\
&\quad\le\sup_{r,\hat{r}}\left\{\bar{D}(r\,\|\,\hat{r}):D_{\mathrm{KL}}(r\,\|\,\hat{r})\le\epsilon\right\} \\
&\qquad+\Pr\big(D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))>\epsilon\big)\mathbb{E}\left[\bar{D}(r(X)\,\|\,\hat{r}(X))\,\middle|\,D_{\mathrm{KL}}(r(X)\,\|\,\hat{r}(X))>\epsilon\right],
\end{aligned}$$

where we have bounded the first expectation by the conditional supremum. Applying Lemma 24 to the second expectation,

$$\begin{aligned}
&\mathbb{E}\left[\bar{D}(r(X) \,\|\, \hat{r}(X))\right] \\
&\quad \le \sup_{r,\hat{r}} \left\{\bar{D}(r \,\|\, \hat{r}) : D_{\mathrm{KL}}(r \,\|\, \hat{r}) \le \epsilon\right\} + 2\Pr\big(D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)) > \epsilon\big) \\
&\qquad + \Pr\big(D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)) > \epsilon\big)\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)) \mid D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)) > \epsilon\right] \\
&\quad \le \sup_{r,\hat{r}} \left\{\bar{D}(r \,\|\, \hat{r}) : D_{\mathrm{KL}}(r \,\|\, \hat{r}) \le \epsilon\right\} + 2\Pr\big(D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X)) > \epsilon\big) \\
&\qquad + \mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X))\right] \\
&\quad \le \sup_{r,\hat{r}} \left\{\bar{D}(r \,\|\, \hat{r}) : D_{\mathrm{KL}}(r \,\|\, \hat{r}) \le \epsilon\right\} + \left(1 + \frac{2}{\epsilon}\right)\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X))\right]. \qquad (62)
\end{aligned}$$

The second inequality above is implied by a similar application of total expectation to $\mathbb{E}\left[D_{\mathrm{KL}}(r(X) \,\|\, \hat{r}(X))\right]$, and the third inequality is due to Markov's inequality. By Assumption 6, the last term in (62) converges to zero in probability for any $\epsilon > 0$.

We now take $\epsilon \to 0$ and argue that the first right-hand side term in (62) also converges to zero. This is because the condition $D_{\mathrm{KL}}(r \,\|\, \hat{r}) \le \epsilon$ excludes the cases $\hat{r} = 0$ unless $r = 0$, and $\hat{r} = 1$ unless $r = 1$, which would cause $\bar{D}(r \,\|\, \hat{r})$ to diverge. $\bar{D}(r \,\|\, \hat{r})$ is therefore bounded and continuous on the set $\{(r, \hat{r}) : D_{\mathrm{KL}}(r \,\|\, \hat{r}) \le \epsilon\}$. As $\epsilon \to 0$, this set shrinks toward the line $r = \hat{r}$ where $\bar{D}(r \,\|\, \hat{r}) = 0$. The lemma is thus proven by combining (61), (62) and taking $\epsilon \to 0$. ■

### A.5.2 Auxiliary Lemmas for Lemma 16

**Lemma 22** *For $r^*(\mu; r)$ defined in (10),*

$$\begin{aligned}
\sup_{\mu} \left|\log \frac{r^*(\mu; r)}{r^*(\mu; \hat{r})}\right| &= \left|\log \frac{1 - \sqrt{1 - r}}{1 - \sqrt{1 - \hat{r}}}\right|, \\
\sup_{\mu} \left|\log \frac{1 - r^*(\mu; r)}{1 - r^*(\mu; \hat{r})}\right| &= \left|\log \frac{1 - \sqrt{r}}{1 - \sqrt{\hat{r}}}\right|.
\end{aligned}$$

**Proof** We prove the first identity. From (10) we have

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log r^*(\mu; r) &= \frac{1}{1 + \mu - \sqrt{(1+\mu)^2 - 4r\mu}} \left(1 - \frac{1 + \mu - 2r}{\sqrt{(1+\mu)^2 - 4r\mu}}\right) - \frac{1}{\mu} \\
&= \frac{\sqrt{(1+\mu)^2 - 4r\mu} - (1+\mu) + 2r}{1 + \mu - \sqrt{(1+\mu)^2 - 4r\mu}} \frac{1}{\sqrt{(1+\mu)^2 - 4r\mu}} - \frac{1}{\mu} \\
&= \frac{-4r\mu + 2r\left(1 + \mu + \sqrt{(1+\mu)^2 - 4r\mu}\right)}{4r\mu\sqrt{(1+\mu)^2 - 4r\mu}} - \frac{1}{\mu} \\
&= \frac{1 - \mu}{2\mu\sqrt{(1+\mu)^2 - 4r\mu}} - \frac{1}{2\mu},
\end{aligned}$$

where the third equality comes from multiplying numerator and denominator by $1+\mu+\sqrt{(1+\mu)^2-4r\mu}$ and using the identity $(a-b)(a+b)=a^2-b^2$. Hence

$$\begin{aligned}\frac{\partial}{\partial\mu}\log\frac{r^*(\mu;r)}{r^*(\mu;\hat{r})} &= \frac{1-\mu}{2\mu}\left(\frac{1}{\sqrt{(1+\mu)^2-4r\mu}}-\frac{1}{\sqrt{(1+\mu)^2-4\hat{r}\mu}}\right)\\ &= \frac{1-\mu}{2\mu}\left(\frac{\sqrt{(1+\mu)^2-4\hat{r}\mu}-\sqrt{(1+\mu)^2-4r\mu}}{\sqrt{(1+\mu)^2-4r\mu}\sqrt{(1+\mu)^2-4\hat{r}\mu}}\right)\\ &= \frac{(1-\mu)\big(r^*(\mu;r)-r^*(\mu;\hat{r})\big)}{\sqrt{(1+\mu)^2-4r\mu}\sqrt{(1+\mu)^2-4\hat{r}\mu}},\end{aligned} \tag{63}$$

using the definition of $r^*(\mu;r)$ (10) in the last line.

We now consider three cases: (1) $r=\hat{r}$, (2) $r>\hat{r}$, and (3) $r<\hat{r}$. (1) If $r=\hat{r}$, then $\log(r^*(\mu;r)/r^*(\mu;\hat{r}))=0$ for all $\mu$ and the identity is true. (2) For $r>\hat{r}$, Lemma 2 implies that $r^*(\mu;r)>r^*(\mu;\hat{r})$ also. It follows that $\log(r^*(\mu;r)/r^*(\mu;\hat{r}))$ is positive for all $\mu$. Furthermore from (63), $\log(r^*(\mu;r)/r^*(\mu;\hat{r}))$ increases with $\mu$ for $\mu<1$ and decreases for $\mu>1$. The maximum therefore occurs at $\mu=1$, and the substitution of $r^*(1;r)=1-\sqrt{1-r}$ yields the desired identity. (3) For $r<\hat{r}$, the same arguments show that $\log(r^*(\mu;r)/r^*(\mu;\hat{r}))$ is negative for all $\mu$, decreases with $\mu$ for $\mu<1$, and increases for $\mu>1$. The maximum absolute value occurs therefore at $\mu=1$ as well.

The proof of the second identity in the lemma statement is analogous by symmetry and shows that the maximizing value is $\mu=-1$. ■

**Lemma 23** *For $r,\hat{r}\in[0,1]$,*

$$\begin{aligned}-0.6140 &\le r\log\frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}\le r\log\frac{r}{\hat{r}}+\log 2,\\ -0.6140 &\le (1-r)\log\frac{1-\sqrt{r}}{1-\sqrt{\hat{r}}}\le (1-r)\log\frac{1-r}{1-\hat{r}}+\log 2.\end{aligned}$$

**Proof** As with the proof of Lemma 22, we prove only the first line of inequalities. The second line follows by symmetry.

To obtain the lower bound, we observe that the quantity of interest is minimized for any $r\in[0,1]$ by taking $\hat{r}=1$. Numerical minimization of the resulting quantity $r\log(1-\sqrt{1-r})$ over $r\in[0,1]$ then yields the lower bound.

To obtain the upper bound, we maximize the quantity

$$r\log\frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}-r\log\frac{r}{\hat{r}}=r\log\left(\frac{1-\sqrt{1-r}}{r}\frac{\hat{r}}{1-\sqrt{1-\hat{r}}}\right). \tag{64}$$

It can be verified that

$$\frac{d}{d\hat{r}}\frac{\hat{r}}{1-\sqrt{1-\hat{r}}}=-\frac{1}{2\sqrt{1-\hat{r}}}<0, \tag{65}$$

which implies that (64) is monotonically decreasing in $\hat{r}$ for any $r\in[0,1]$ and is maximized by taking $\hat{r}\to 0$. By l'Hôpital's rule,

$$\lim_{\hat{r}\to 0}\frac{\hat{r}}{1-\sqrt{1-\hat{r}}}=2, \tag{66}$$

and it remains to maximize

$$r \log \frac{2(1-\sqrt{1-r})}{r}. \tag{67}$$

The calculation in (65) also implies that $\log(2(1-\sqrt{1-r})/r)$ is monotonically *increasing* in $r$, and (66) implies that

$$\lim_{r\to 0} \log \frac{2(1-\sqrt{1-r})}{r} = 0,$$

so that $\log(2(1-\sqrt{1-r})/r) \geq 0$ for $r \in [0,1]$. It follows that the quantity in (67) is monotonically increasing in $r$, as the product of two non-negative and monotonically increasing functions, and is therefore maximized at $r = 1$, yielding $\log 2$. This proves the upper bound. ■

**Lemma 24** *For $r, \hat{r} \in [0,1]$,*

$$\bar{D}(r \parallel \hat{r}) \equiv r\left|\log \frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}\right| + (1-r)\left|\log \frac{1-\sqrt{r}}{1-\sqrt{\hat{r}}}\right| \leq D_{\mathrm{KL}}(r \parallel \hat{r}) + 2.$$

**Proof** Again it suffices to bound the first term of $\bar{D}(r \parallel \hat{r})$ because the second term is analogous. From Lemma 23 we have

$$\begin{aligned} r\left|\log \frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}\right| &= \max\left\{r \log \frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}, -r \log \frac{1-\sqrt{1-r}}{1-\sqrt{1-\hat{r}}}\right\} \\ &\leq \max\left\{r \log \frac{r}{\hat{r}} + \log 2, 0.6140\right\}, \end{aligned}$$

where we note that $r \log(r/\hat{r}) + \log 2$ is always positive since its minimum value is $\log 2 - 1/e = 0.3253$ at $(r, \hat{r}) = (1/e, 1)$. We may further and more simply bound the above by

$$\max\left\{r \log \frac{r}{\hat{r}} + \log 2, 0.6140\right\} \leq r \log \frac{r}{\hat{r}} + \frac{1}{e} + 0.6140 < \log \frac{r}{\hat{r}} + 1,$$

from which the result follows. ■

### A.5.3 Proof of Lemma 17

**Proof** Let $\hat{\mu} = \hat{\lambda}^T \hat{\mathbf{f}}(a, \hat{r}(x))$ and $\mu = \hat{\lambda}^T \mathbf{f}(a, r(x))$. By the mean value theorem,

$$|H_b\left(r(x), r^*(\hat{\mu}; r(x))\right) - H_b\left(r(x), r^*(\mu; r(x))\right)| = \left|\frac{\partial H_b\big(r, r^*(\mu; r)\big)}{\partial \mu}\right|_{\mu=\hat{\lambda}^T \bar{\mathbf{f}}} \Bigg| |\hat{\mu} - \mu|,$$

where $\bar{\mathbf{f}}$ is a convex combination of $\hat{\mathbf{f}}(a, \hat{r}(x))$ and $\mathbf{f}(a, r(x))$. By differentiating (12) with respect to $\mu$ and combining with (90), we find that

$$\left|\frac{\partial H_b\big(r, r^*(\mu; r)\big)}{\partial \mu}\right| = |\mu| \left|\frac{\partial r^*(\mu; r)}{\partial \mu}\right|,$$

and it can be verified using (91) that $|\partial r^*(\mu; r)/\partial \mu| \leq 1$. Hence

$$\begin{aligned}|H_b\left(r(x), r^*(\hat{\mu}; r(x))\right) - H_b\left(r(x), r^*(\mu; r(x))\right)| &\leq \left|\hat{\lambda}^T \bar{\mathbf{f}}\right| \left|\hat{\lambda}^T \left(\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right)\right| \\ &\leq \left\|\hat{\lambda}\right\|_1^2 \left\|\bar{\mathbf{f}}\right\|_\infty \left\|\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right\|_\infty \\ &\leq \left(\frac{\log 2}{\epsilon}\right)^2 \left\|\bar{\mathbf{f}}\right\|_\infty \left\|\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right\|_\infty,\end{aligned} \tag{68}$$

where the second line results from two applications of Hölder's inequality, and the third line from Lemma 9 given $\hat{\lambda} \in \Lambda_0$. Since Lemma 21 applies to both $\hat{\mathbf{f}}(a, \hat{r}(x))$ and $\mathbf{f}(a, r(x))$, we have

$$\left\|\bar{\mathbf{f}}\right\|_\infty \leq \frac{1}{\eta} - 1 \tag{69}$$

for their convex combination as well. Using the triangle inequality,

$$\left\|\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right\|_\infty \leq \left\|\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, \hat{r}(x))\right\|_\infty + \left\|\mathbf{f}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right\|_\infty. \tag{70}$$

In the case of MSP, the second right-hand side term is zero because $\mathbf{f}$ does not depend on $r$. For GEO, (25) implies that

$$\begin{aligned}&\left\|\mathbf{f}(a, \hat{r}(x)) - \mathbf{f}(a, r(x))\right\|_\infty \\ &\qquad \leq |\hat{r}(x) - r(x)| \max_{a \in \mathcal{A}, y \in \{0,1\}} \max\left\{\frac{1}{p_Y(y)}\left(\frac{1}{p_{A\,|\,Y}(a\,|\,y)} - 1\right), \frac{1}{p_Y(y)}\right\} \\ &\qquad \leq |\hat{r}(x) - r(x)| \left(\frac{1}{\eta} - 1\right),\end{aligned} \tag{71}$$

where the last inequality was derived in the proof of Lemma 21.

We combine (68), (69), (70), and (71) to obtain

$$\begin{aligned}&|H_b\left(r(x), r^*(\hat{\mu}; r(x))\right) - H_b\left(r(x), r^*(\mu; r(x))\right)| \\ &\qquad \leq \left(\frac{\log 2}{\epsilon}\right)^2 \left(\frac{1}{\eta} - 1\right) \left(\left\|\hat{\mathbf{f}}(a, \hat{r}(x)) - \mathbf{f}(a, \hat{r}(x))\right\|_\infty + \left(\frac{1}{\eta} - 1\right) |\hat{r}(x) - r(x)|\right),\end{aligned}$$

where the $|\hat{r}(x) - r(x)|$ term is absent in the MSP case. Taking expectations over $x \in \mathcal{X}$,

$$\begin{aligned}&\left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T \hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T \mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right| \\ &\quad \leq \mathbb{E}\left[\left|H_b\left(r(X), r^*\left(\hat{\lambda}^T \hat{\mathbf{f}}(A, \hat{r}(X)); r(X)\right)\right) - H_b\left(r(X), r^*\left(\hat{\lambda}^T \mathbf{f}(A, r(X)); r(X)\right)\right)\right|\right] \\ &\quad \leq \left(\frac{\log 2}{\epsilon}\right)^2 \left(\frac{1}{\eta} - 1\right) \left(\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A, \hat{r}(X)) - \mathbf{f}(A, \hat{r}(X))\right\|_\infty\right] + \left(\frac{1}{\eta} - 1\right) \mathbb{E}\left[|\hat{r}(X) - r(X)|\right]\right).\end{aligned}$$

The proof of Lemma 12 shows that $\mathbb{E}\left[\left\|\hat{\mathbf{f}}(A, \hat{r}(X)) - \mathbf{f}(A, \hat{r}(X))\right\|_\infty\right]$ converges to zero in probability as $\hat{p}_A$, $\hat{p}_{A,Y}$, $\hat{p}_Y$ converge to the true probabilities. Assumption 8 ensures that $\mathbb{E}\left[|\hat{r}(X) - r(X)|\right]$ converges to zero in probability as well, completing the proof. ■

### A.5.4 Proof of Lemma 18

**Proof** We apply the mean value theorem in the same way as in the proof of Lemma 17, where now $\hat{\mu} = \hat{\lambda}^T \mathbf{f}(a, r(x))$, $\mu = \lambda^{*T} \mathbf{f}(a, r(x))$, and the intermediate value is $\bar{\mu} = \bar{\lambda}^T \mathbf{f}(a, r(x))$ for some convex combination $\bar{\lambda}$ of $\hat{\lambda}$ and $\lambda^*$. Following the same steps as in the earlier proof, we obtain

$$\begin{aligned}
|H_b\left(r(x), r^*(\hat{\mu}; r(x))\right) - H_b\left(r(x), r^*(\mu; r(x))\right)| &\leq \left|\bar{\lambda}^T \mathbf{f}\right| \left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(a, r(x))\right| \\
&\leq \left\|\bar{\lambda}\right\|_1 \|\mathbf{f}\|_\infty \left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(a, r(x))\right| \\
&\leq \left(\frac{\log 2}{\epsilon}\right)\left(\frac{1}{\eta} - 1\right) \left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(a, r(x))\right|,
\end{aligned}$$

again using Hölder's inequality in the second line, and Lemmas 9 and 21 in the third line. Taking expectations over $x \in \mathcal{X}$ then yields

$$\begin{aligned}
&\left|\mathbb{E}\left[H_b\left(r(X), r^*\left(\hat{\lambda}^T \mathbf{f}(A, r(X)); r(X)\right)\right)\right] - \mathbb{E}\left[H_b\left(r(X), r^*\left(\lambda^{*T} \mathbf{f}(A, r(X)); r(X)\right)\right)\right]\right| \\
&\quad \leq \mathbb{E}\left[\left|H_b\left(r(X), r^*\left(\hat{\lambda}^T \mathbf{f}(A, r(X)); r(X)\right)\right) - H_b\left(r(X), r^*\left(\lambda^{*T} \mathbf{f}(A, r(X)); r(X)\right)\right)\right|\right] \\
&\quad \leq \left(\frac{\log 2}{\epsilon}\right)\left(\frac{1}{\eta} - 1\right) \mathbb{E}\left[\left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(A, r(X))\right|\right].
\end{aligned}$$

The proof is completed by Lemma 25 below. ■

**Lemma 25** *Under Assumptions 2, 3, 4, 7, 9,*

$$\mathbb{E}\left[\left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(A, r(X))\right|\right] \xrightarrow{p} 0.$$

**Proof** We prove that the quantity in question converges to zero in the $L_2$ norm,

$$\mathbb{E}\left[\left|\left(\hat{\lambda} - \lambda^*\right)^T \mathbf{f}(A, r(X))\right|^2\right] \xrightarrow{p} 0, \tag{72}$$

As shown in the proof of Lemma 8, this implies convergence in the $L_1$ norm as in the lemma statement.

To establish (72), we use Theorem 6, which implies that for any $\varepsilon > 0$, we have $J(\hat{\lambda}) \leq J(\lambda^*) + \varepsilon$ with probability converging to 1 as $n, m \to \infty$. Assume then that $J(\hat{\lambda}) \leq J(\lambda^*) + \varepsilon$. Define $\lambda_\alpha = \alpha\lambda^* + (1 - \alpha)\hat{\lambda}$ for $\alpha \in [0, 1]$, and

$$G(\lambda) = \mathbb{E}\left[g\left(\lambda^T \mathbf{f}(A, r(X)); r(X)\right)\right] \tag{73}$$

to be the first term in the population dual objective function (14), (15). From Bertsekas (1999, Proposition A.23b), we have the following second-order expansion:

$$\begin{aligned}
G\left(\hat{\lambda}\right) &= G(\lambda_\alpha) + \left(\hat{\lambda} - \lambda_\alpha\right)^T \nabla G(\lambda_\alpha) + \frac{1}{2}\left(\hat{\lambda} - \lambda_\alpha\right)^T \nabla^2 G(\bar{\lambda})\left(\hat{\lambda} - \lambda_\alpha\right) \\
&= G(\lambda_\alpha) + \alpha\left(\hat{\lambda} - \lambda^*\right)^T \nabla G(\lambda_\alpha) + \frac{1}{2}\alpha^2\left(\hat{\lambda} - \lambda^*\right)^T \nabla^2 G(\bar{\lambda})\left(\hat{\lambda} - \lambda^*\right),
\end{aligned} \tag{74}$$

where $\bar{\lambda}$ is some point on the line segment between $\hat{\lambda}$ and $\lambda_\alpha$. By differentiating (73) and using (90), we find

$$\begin{aligned}\nabla G(\lambda) &= \mathbb{E}\left[-r^*\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\mathbf{f}(A,r(X))\right],\\ \nabla^2 G(\lambda) &= \mathbb{E}\left[s\big(\lambda^T\mathbf{f}(A,r(X));r(X)\big)\mathbf{f}(A,r(X))\mathbf{f}^T(A,r(X))\right],\end{aligned} \tag{75}$$

where $s(\mu;r) = -\partial r^*(\mu;r)/\partial\mu$. Substituting (75) into (74) yields

$$G\big(\hat{\lambda}\big) = G(\lambda_\alpha)+\alpha\big(\hat{\lambda}-\lambda^*\big)^T\nabla G(\lambda_\alpha)+\frac{1}{2}\alpha^2\mathbb{E}\left[s\big(\bar{\lambda}^T\mathbf{f}(A,r(X));r(X)\big)\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right].$$

Since $\bar{\lambda}$ also lies on the line segment between $\hat{\lambda}$ and $\lambda^*$, the application of Assumption 9 yields

$$G\big(\hat{\lambda}\big) \geq G(\lambda_\alpha) + \alpha\big(\hat{\lambda}-\lambda^*\big)^T\nabla G(\lambda_\alpha) + \frac{1}{2}\alpha^2\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right]. \tag{76}$$

By repeating the above steps for $\lambda^*$ in place of $\hat{\lambda}$, we also obtain

$$G\big(\lambda^*\big) \geq G(\lambda_\alpha) + (1-\alpha)\big(\lambda^*-\hat{\lambda}\big)^T\nabla G(\lambda_\alpha) + \frac{1}{2}(1-\alpha)^2\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right]. \tag{77}$$

Multiplying (76) by $1-\alpha$, (77) by $\alpha$, and summing,

$$(1-\alpha)G\big(\hat{\lambda}\big) + \alpha G\big(\lambda^*\big) \geq G(\lambda_\alpha) + \frac{1}{2}\alpha(1-\alpha)\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right]. \tag{78}$$

Since the $\ell_1$ norm is convex, we also have

$$(1-\alpha)\epsilon\big\|\hat{\lambda}\big\|_1 + \alpha\epsilon\big\|\lambda^*\big\|_1 \geq \epsilon\big\|\lambda_\alpha\big\|_1. \tag{79}$$

Adding (78) and (79),

$$\begin{aligned}(1-\alpha)J\big(\hat{\lambda}\big) + \alpha J\big(\lambda^*\big) &\geq J(\lambda_\alpha) + \frac{1}{2}\alpha(1-\alpha)\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right]\\ &\geq J\big(\lambda^*\big) + \frac{1}{2}\alpha(1-\alpha)\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right],\end{aligned}$$

where the last inequality is due to the optimality of $\lambda^*$. Using the assumption $J(\hat{\lambda}) \leq J(\lambda^*) + \varepsilon$, we arrive at

$$\frac{1}{2}\alpha\tau\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right] \leq \varepsilon,$$

and since $\alpha \in [0,1]$ was arbitrary, we take $\alpha \to 1$ to yield

$$\mathbb{E}\left[\left|\big(\hat{\lambda}-\lambda^*\big)^T\mathbf{f}(A,r(X))\right|^2\right] \leq \frac{2\varepsilon}{\tau}. \tag{80}$$

We have thus shown that the near-optimality condition $J(\hat{\lambda}) \leq J(\lambda^*) + \varepsilon$ implies (80) for any $\varepsilon > 0$. Since the former holds with probability converging to 1, this proves (72).

■

## Appendix B. ADMM Algorithm Details

In this appendix, we present a closed-form solution for step (20a) in the ADMM algorithm of Section 4.2. We then describe alternative ADMM algorithms in Appendix B.2.

### B.1 Closed-Form Solution for (20a)

To simplify notation, define $\tilde{\rho} = \rho n$ and drop the index $i$ and the hat from $\hat{r}$. Then the first-order optimality condition for (20a) can be written as

$$n\frac{\partial \mathsf{obj}(\mu)}{\partial \mu} = \tilde{\rho}(\mu - b) - r^*(\mu; r) = 0. \tag{81}$$

Figure 1a shows that $r^*(\mu; r)$ decreases monotonically with $\mu$, and this can be proven by showing that expression (91) for $dr^*(\mu; r)/d\mu$ does not change sign. It follows that the quantity in (81) strictly increases with $\mu$ and the equation has a unique solution.

To solve for the root of (81), we use (10) and rearrange to isolate the square root on one side:

$$1 + \mu - 2\tilde{\rho}\mu(\mu - b) = \sqrt{(1+\mu)^2 - 4r\mu}.$$

Upon squaring both sides, it is seen that the zeroth-order terms in $\mu$ cancel to give

$$\tilde{\rho}^2\mu^2(\mu - b)^2 - \tilde{\rho}\mu(1+\mu)(\mu - b) = -r\mu. \tag{82}$$

One solution to (82) is $\mu = 0$ but it satisfies the original condition (81) only if $r = -\tilde{\rho}b$. Assuming this is not the case, we divide both sides of (82) by $\mu$ and expand to yield the following cubic equation:

$$\mu^3 \underbrace{-(2b + (1/\tilde{\rho}))}_{a_2}\mu^2 + \underbrace{(b^2 + (b-1)/\tilde{\rho})}_{a_1}\mu + \underbrace{b/\tilde{\rho} + r/\tilde{\rho}^2}_{a_0} = 0. \tag{83}$$

Toward obtaining a closed-form expression for the roots of (83), define the coefficients of the corresponding *depressed cubic* equation as

$$\begin{aligned} p &= 3a_1 - a_2^2, \\ q &= a_2^3 - \frac{9}{2}a_1a_2 + \frac{27}{2}a_0. \end{aligned}$$

We find empirically that cubic equation (83) always has three real roots and that the desired root corresponding to the solution of (81) is given by the same one of these roots. We do not however have proofs of these facts. In the case of three real roots, they are given by the trigonometric formula

$$\mu^* = \frac{1}{3}\left(2\sqrt{-p}\cos\left(\frac{1}{3}\left(\arccos\left(\frac{q}{p\sqrt{-p}}\right) - 2k\pi\right)\right) - a_2\right), \quad k = 0, 1, 2, \tag{84}$$

and the desired root appears to always correspond to $k = 1$. In any case, the correct root can be identified by first noting that since $r^*(\mu; r) \in [0, 1]$, the solution to (81) must lie in the interval $[b, b + (1/\tilde{\rho})]$. We may then compute all three roots in (84) and choose the one lying in $[b, b + (1/\tilde{\rho})]$.

### B.2 Alternative ADMM Algorithms

This section presents alternative ADMM decompositions for the dual problems corresponding to MSP (14) and GEO (15). To simplify notation, we suppress the hat symbols on $\hat{r}(x)$ and $\hat{\mathbf{f}}(x)$.

#### B.2.1 Mean Score Parity

Define auxiliary variables $\tilde{\lambda}_a$ as follows:

$$\tilde{\lambda}_a = \frac{\lambda_a}{p_A(a)} - \sum_{a' \in \mathcal{A}} \lambda_{a'}, \quad a \in \mathcal{A}, \tag{85}$$

with $\tilde{\lambda} = (\tilde{\lambda}_a)_{a \in \mathcal{A}}$. Then the empirical version of (14) can be written as

$$\begin{aligned} \min_{\lambda, \tilde{\lambda}} \quad & \frac{1}{n} \sum_{i=1}^{n} g\big(\mu_i; r_i\big) + \epsilon \|\lambda\|_1 \\ \text{s. t.} \quad & \mu_i = \sum_{a \in \mathcal{A}} p_{A \,|\, X}(a \,|\, x_i) \tilde{\lambda}_a, \end{aligned} \tag{86}$$

where $\mu_i = \mu(x_i)$, $r_i = r(x_i)$, and we regard $\lambda$ and $\tilde{\lambda}$ as two sets of optimization variables that are linearly related through (85). Let $\mathbf{B} \in \mathbb{R}^{n \times d}$ be a matrix with entries $\mathbf{B}_{i,a} = p_{A \,|\, X}(a \,|\, x_i)$ and rows $\mathbf{b}_i^T$ so that we may write $\mu = \mathbf{B}\tilde{\lambda}$, $\mu_i = \mathbf{b}_i^T \tilde{\lambda}$. The objective function in (86) is therefore separable between $\lambda$ and $\tilde{\lambda}$. With $\mathbf{1}$ denoting a vector of ones and $\mathbf{P}_A$ the $d \times d$ diagonal matrix with diagonal entries $p_A(a)$, a scaled ADMM algorithm for (86) consists of the following three steps in each iteration $k = 0, 1, \ldots$:

$$\tilde{\lambda}^{k+1} = \arg\min_{\tilde{\lambda}} \; \frac{1}{n} \sum_{i=1}^{n} g\big(\mathbf{b}_i^T \tilde{\lambda}; r_i\big) + \frac{\rho}{2} \left\| \tilde{\lambda} - \left(\mathbf{P}_A^{-1} - \mathbf{1}\mathbf{1}^T\right) \lambda^k + u^k \right\|_2^2 \tag{87}$$

$$\lambda^{k+1} = \arg\min_{\lambda} \; \epsilon \|\lambda\|_1 + \frac{\rho}{2} \left\| \left(\mathbf{P}_A^{-1} - \mathbf{1}\mathbf{1}^T\right) \lambda - \tilde{\lambda}^{k+1} - u^k \right\|_2^2 \tag{88}$$

$$u^{k+1} = u^k + \tilde{\lambda}^{k+1} - \left(\mathbf{P}_A^{-1} - \mathbf{1}\mathbf{1}^T\right) \lambda^{k+1}. \tag{89}$$

The optimization in (88) is an $\ell_1$-penalized quadratic minimization and can be handled by many convex solvers. The optimization in (87) can be solved using Newton's method. Below we give the gradient and Hessian of the first term in (87); the second Euclidean norm term is standard. First, using the definition of $g(\mu; r)$ in (12), we find that

$$\frac{dg(\mu; r)}{d\mu} = -r^*(\mu; r) \tag{90}$$

$$\frac{d^2 g(\mu; r)}{d\mu^2} = -\frac{dr^*(\mu; r)}{d\mu} = \begin{cases} \dfrac{1}{2\mu^2} \left(1 - \dfrac{1 + (1-2r)\mu}{\sqrt{(1+\mu)^2 - 4r\mu}}\right), & \mu \neq 0 \\ r(1-r), & \mu = 0. \end{cases} \tag{91}$$

The simple form in (90) is due to $r^*(\mu; r)$ satisfying the optimality condition (33) and the ensuing cancellation of terms. It is also related to Bertsekas (1999, Proposition 6.1.1). The

gradient and Hessian of the first term in (87) are then given by

$$\nabla \left( \frac{1}{n} \sum_{i=1}^{n} g\big(\mathbf{b}_i^T \tilde{\lambda}; r_i\big) \right) = -\frac{1}{n} \mathbf{B}^T \mathbf{r}^* \tag{92}$$

$$\nabla^2 \left( \frac{1}{n} \sum_{i=1}^{n} g\big(\mathbf{b}_i^T \tilde{\lambda}; r_i\big) \right) = -\frac{1}{n} \mathbf{B}^T \mathbf{H} \mathbf{B}, \tag{93}$$

where $\mathbf{r}^*$ is the $n$-dimensional vector with components $r^*(\mu_i; r_i)$ and $\mathbf{H}$ is the $n \times n$ diagonal matrix with entries $dr^*(\mu_i; r_i)/d\mu_i$. In the case where the features $X$ include the protected attribute $A$, $p_{A \mid X}(a \mid x_i) = \mathbf{1}(a = a_i)$, $\mathbf{B}$ is a sparse matrix with a single one in each row, and the Hessian in (93) is diagonal. This implies that optimization (87) is separable over the components of $\tilde{\lambda}$.

### B.2.2 Generalized Equalized Odds

In analogy with (85) we define

$$\tilde{\lambda}_{a,y} = \frac{\lambda_{a,y}}{p_{A \mid Y}(a \mid y)} - \sum_{a' \in \mathcal{A}} \lambda_{a',y}, \quad a \in \mathcal{A},\ y \in \{0, 1\}. \tag{94}$$

Again let $\mathbf{B}$ be a $n \times d$ matrix, recalling that $d = 2|\mathcal{A}|$ in the GEO case, with columns indexed by $(a, y)$ and entries

$$\mathbf{B}_{i,(a,y)} = \begin{cases} \dfrac{(1 - r(x_i)) p_{A \mid X,Y}(a \mid x_i, 0)}{p_Y(0)}, & y = 0 \\ \dfrac{r(x_i) p_{A \mid X,Y}(a \mid x_i, 1)}{p_Y(1)}, & y = 1. \end{cases} \tag{95}$$

It can then be seen from the constraint in (15) that $\mu_i = \mathbf{b}_i^T \tilde{\lambda}$ as before and the empirical version of (15),

$$\min_{\lambda, \tilde{\lambda}} \quad \frac{1}{n} \sum_{i=1}^{n} g\big(\mathbf{b}_i^T \tilde{\lambda}; r_i\big) + \epsilon \|\lambda\|_1, \tag{96}$$

is separable between $\lambda$ and $\tilde{\lambda}$ subject to the linear relation (94). With $\mathbf{P}_{A \mid y}$ for $y = 0, 1$ denoting the $|\mathcal{A}| \times |\mathcal{A}|$ diagonal matrix with diagonal entries $p_{A \mid Y}(a \mid y)$, the three steps in each ADMM iteration for (96) are as follows:

$$\tilde{\lambda}^{k+1} = \arg\min_{\tilde{\lambda}} \ \frac{1}{n} \sum_{i=1}^{n} g\big(\mathbf{b}_i^T \tilde{\lambda}; r_i\big) + \frac{\rho}{2} \sum_{y=0}^{1} \left\| \tilde{\lambda}_{\cdot,y} - \left( \mathbf{P}_{A \mid y}^{-1} - \mathbf{1}\mathbf{1}^T \right) \lambda_{\cdot,y}^k + u_{\cdot,y}^k \right\|_2^2 \tag{97}$$

$$\lambda_{\cdot,y}^{k+1} = \arg\min_{\lambda} \ \epsilon \|\lambda\|_1 + \frac{\rho}{2} \left\| \left( \mathbf{P}_{A \mid y}^{-1} - \mathbf{1}\mathbf{1}^T \right) \lambda - \tilde{\lambda}_{\cdot,y}^{k+1} - u_{\cdot,y}^k \right\|_2^2, \quad y = 0, 1 \tag{98}$$

$$u_{\cdot,y}^{k+1} = u_{\cdot,y}^k + \tilde{\lambda}_{\cdot,y}^{k+1} - \left( \mathbf{P}_{A \mid y}^{-1} - \mathbf{1}\mathbf{1}^T \right) \lambda_{\cdot,y}^{k+1}, \quad y = 0, 1, \tag{99}$$

where $\tilde{\lambda}_{\cdot,y}$, $\lambda_{\cdot,y}$, and $u_{\cdot,y}$ are $|\mathcal{A}|$-dimensional subvectors of $\tilde{\lambda}$, $\lambda$ and $u$ consisting only of components with $y = 0$ or $y = 1$. The optimization in (97) is of the same form as (87) and can also be solved using Newton's method. The same expressions (92), (93) hold for the gradient and Hessian of the first term in (97), where $\mathbf{B}$ is now given by (95). The optimization of $\lambda$ in (98) is separable over $y = 0, 1$ and is the same as step (88) for MSP.

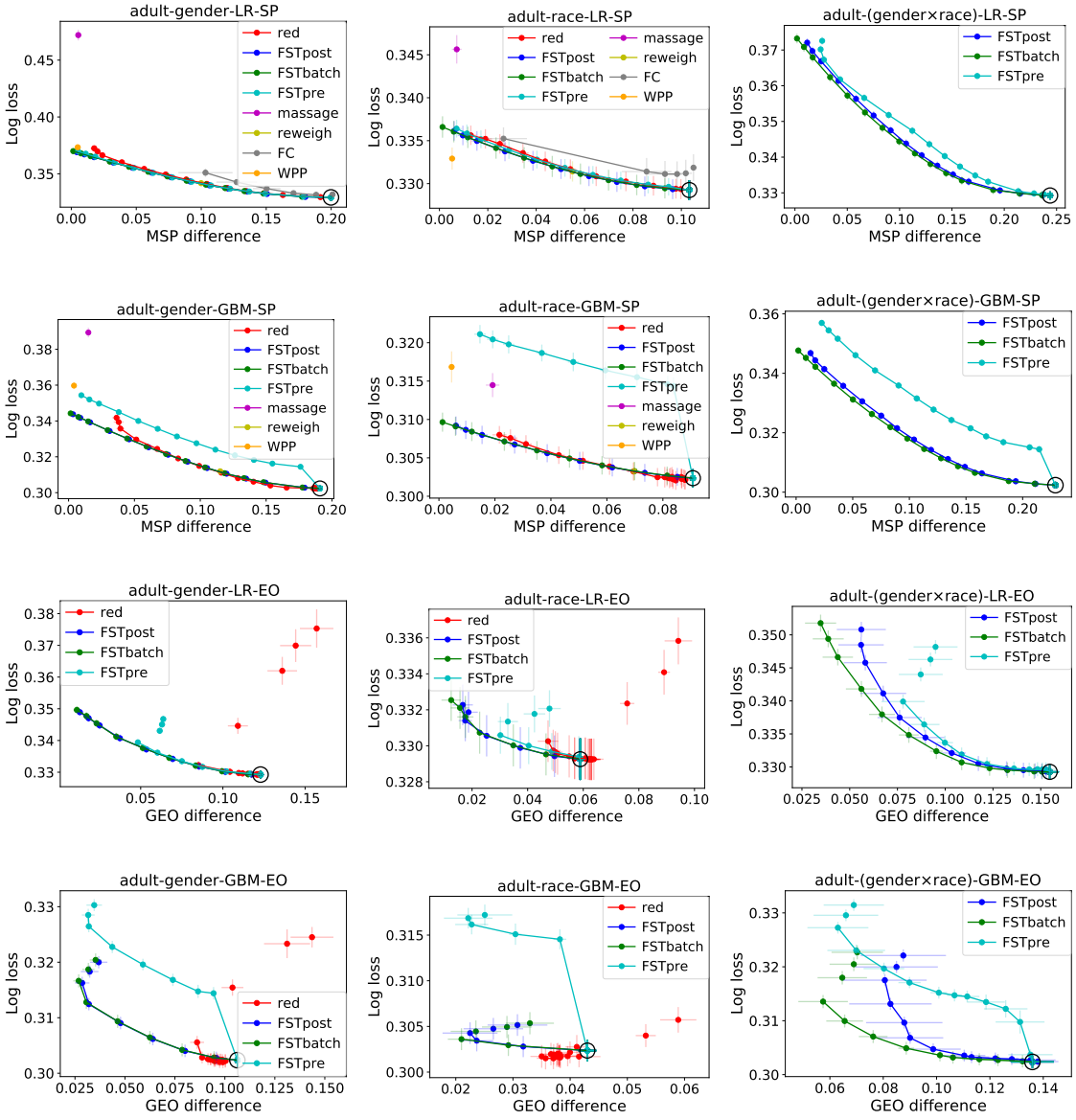



Figure 10: Trade-offs between fairness and log loss on the Adult Income data set with the protected attributes included in the features.

## Appendix C. Additional Experimental Results

This appendix presents results deferred from Section 6, including results with log loss, those for the German credit risk data set, and comparisons with existing methods excluded from the main comparison due to their limitations.

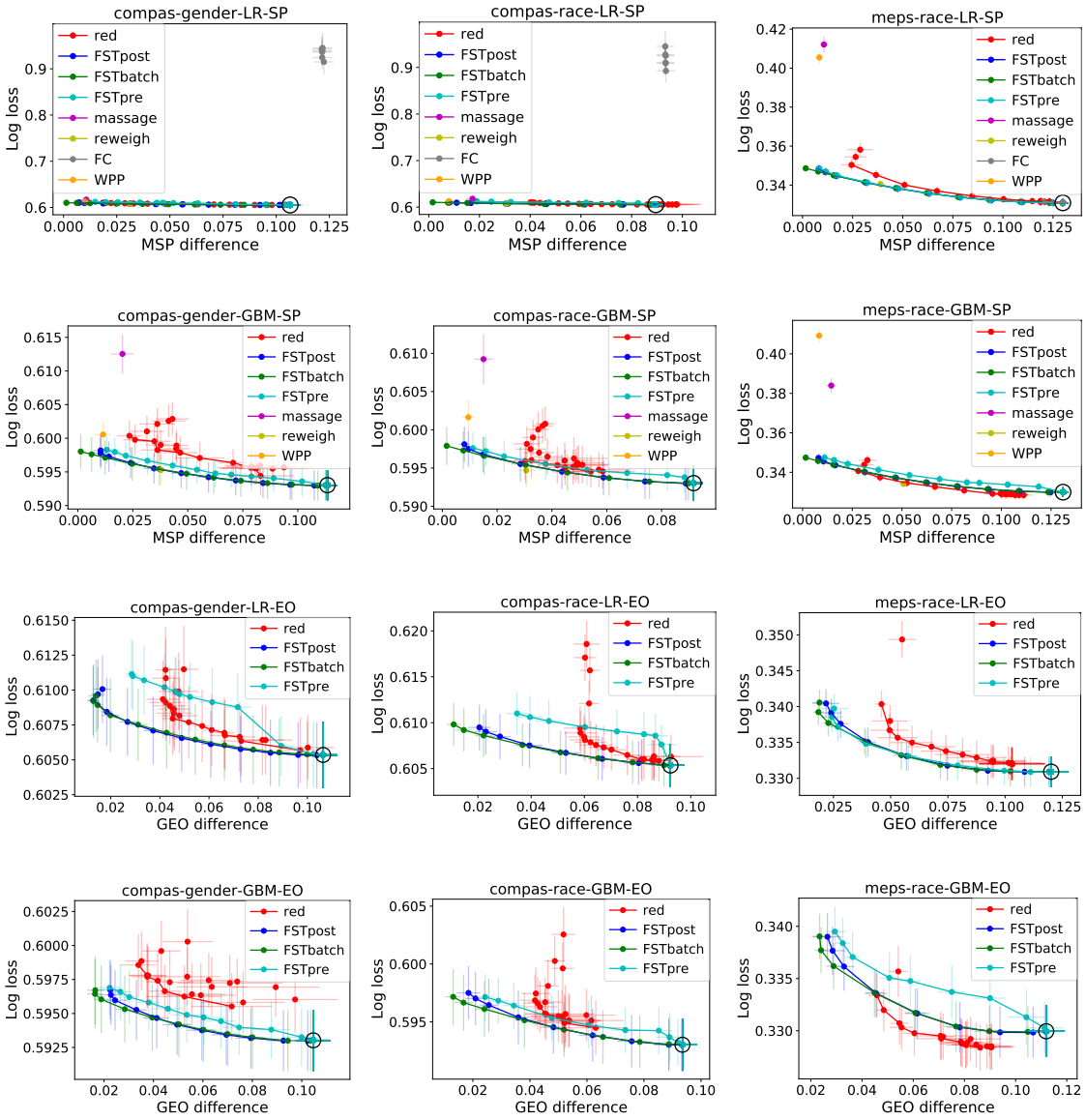



Figure 11: Trade-offs between fairness and log loss on the COMPAS and MEPS data sets with the protected attributes included in the features.

## C.1 Log Loss

Figures 10 and 11 show trade-offs between log loss and MSP or GEO fairness measures for the data set-protected attribute combinations considered in Figures 2–6 and 9 (i.e., with the protected attribute included in the features). The plots are quite similar to those for Brier score in Figures 2–6 and 9.

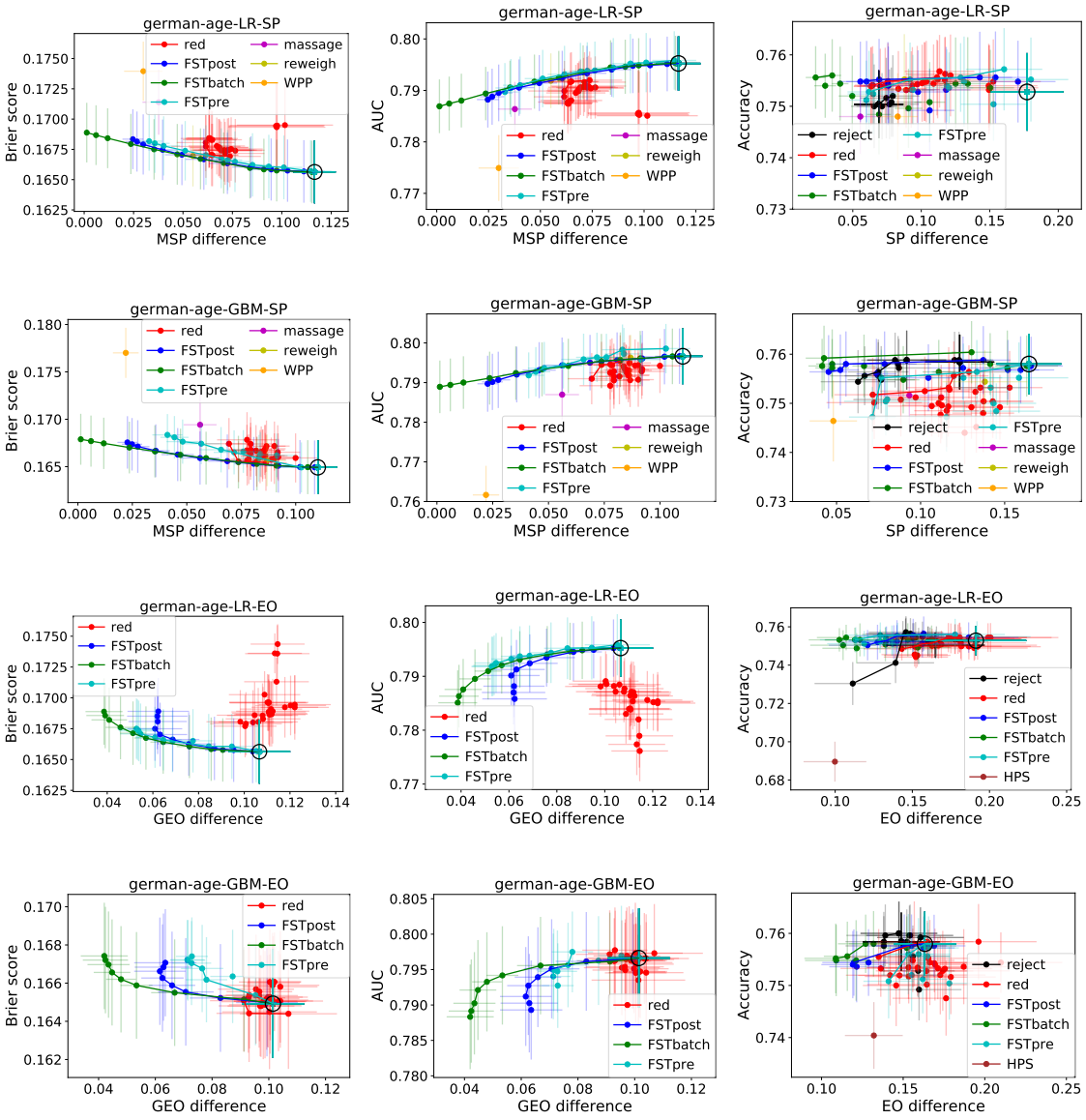



Figure 12: Trade-offs between fairness and classification performance on the German credit data set with age as the protected attribute and the protected attribute included in the features.

## C.2 German Credit Risk Data Set

Figures 12 and 13 depict trade-offs between classification performance and fairness for the German credit data set, where the protected attribute of age is either included in or excluded from the features. These results are included for completeness as German is a standard data

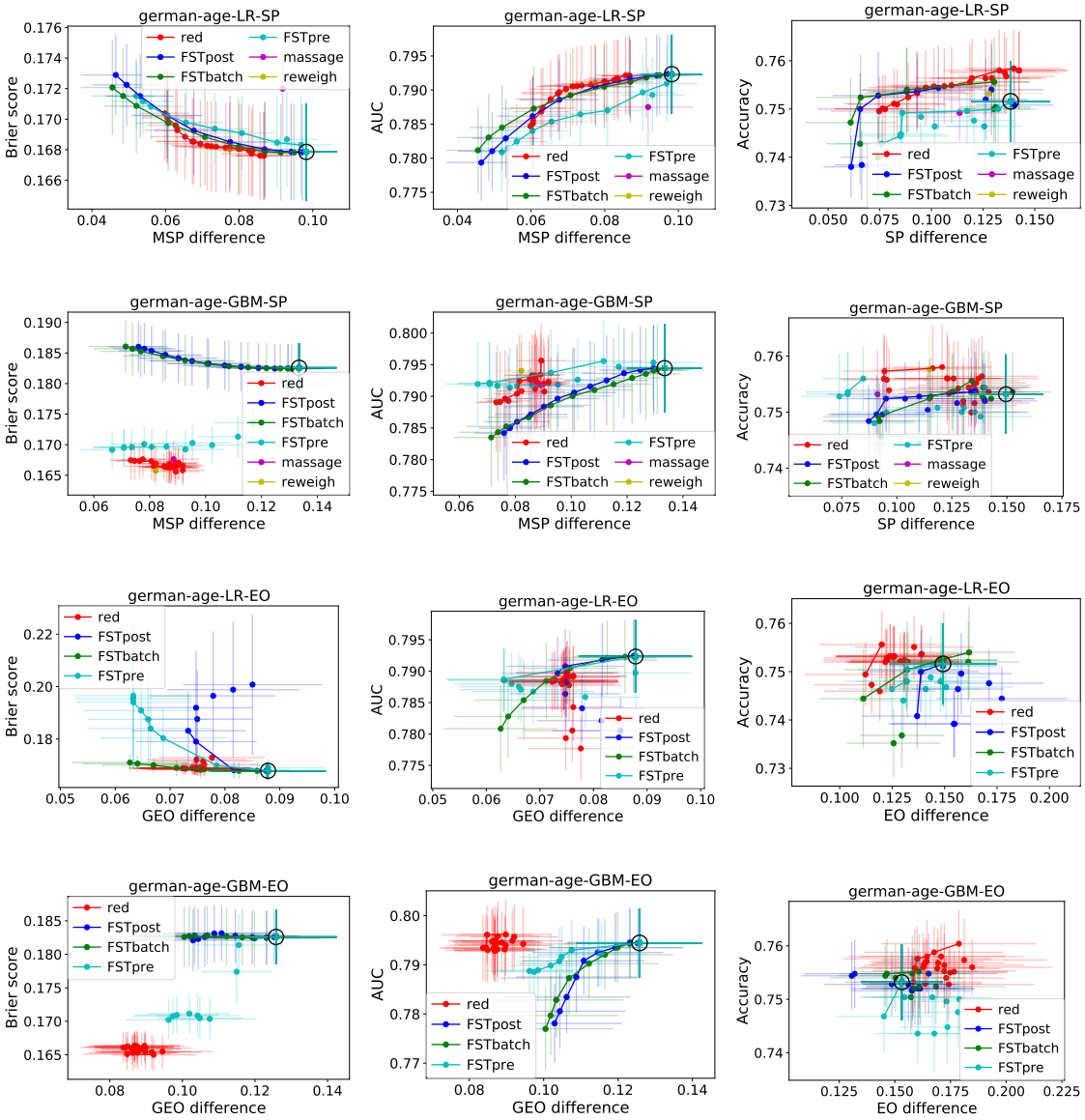



Figure 13: Trade-offs between fairness and classification performance on the German credit data set with age as the protected attribute and the protected attribute excluded from the features.

set in the fairness literature, but the small data set size and consequently large error bars make it difficult to draw conclusions.

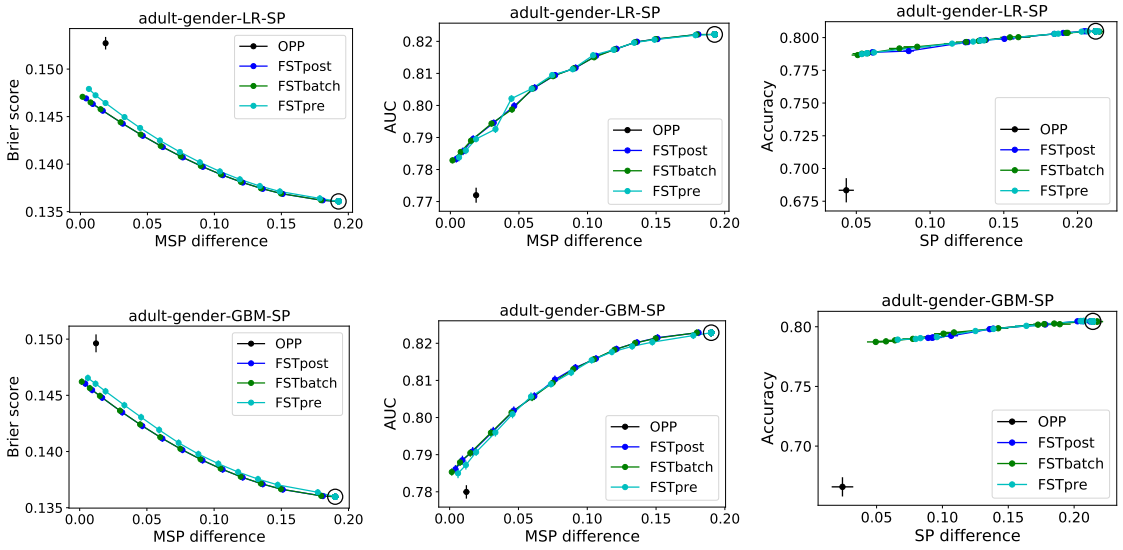



Figure 14: Trade-offs between statistical parity and classification performance measures for the Adult data set with a reduced set of features.

### C.3 Individual Comparisons with Existing Methods

As mentioned in Section 6, we encountered computational difficulties in running the optimized pre-processing (OPP, Calmon et al., 2017) and disparate mistreatment in-processing (DM, Zafar et al., 2017a) methods. In the case of OPP, the method does not scale beyond feature dimensions of $\sim 5$. We have thus conducted separate experiments in which the set of features has been reduced. Figure 14 shows the resulting trade-offs between statistical parity, which is what OPP addresses, and classification performance for the Adult data set. This limited comparison suggests that OPP is not competitive with FST. Unfortunately we were unable to obtain reasonable results for OPP on other data sets so do not show them here.

In the case of DM, when we ran the code[2] on data sets with a full set of features, the optimization either failed to converge or when it did converge, did not appreciably decrease the EO difference from that of an unconstrained logistic regression classifier. (The latter problem was also observed with FC, Zafar et al., 2017c, in Figures 4–6.) We used a constraint type of 4 to impose both FNR and FPR constraints, in keeping with EO, and default values for the disciplined convex-concave programming (DCCP) parameters $\tau$ and $\mu$. For example on the Adult-gender combination, DM failed on 4 of the 10 training folds and converged on the others with little effect, while on the COMPAS-race combination, DM failed on 9 of 10 folds. We noticed that our version of the COMPAS data set has much higher dimension than the one used by Zafar et al. (2017a), due primarily to including a charge description feature and after one-hot encoding of categorical variables. Thus we opted to compare with DM using reduced feature sets, as with OPP. Figure 15 shows the trade-offs obtained on

2. `https://github.com/mbilalzafar/fair-classification`

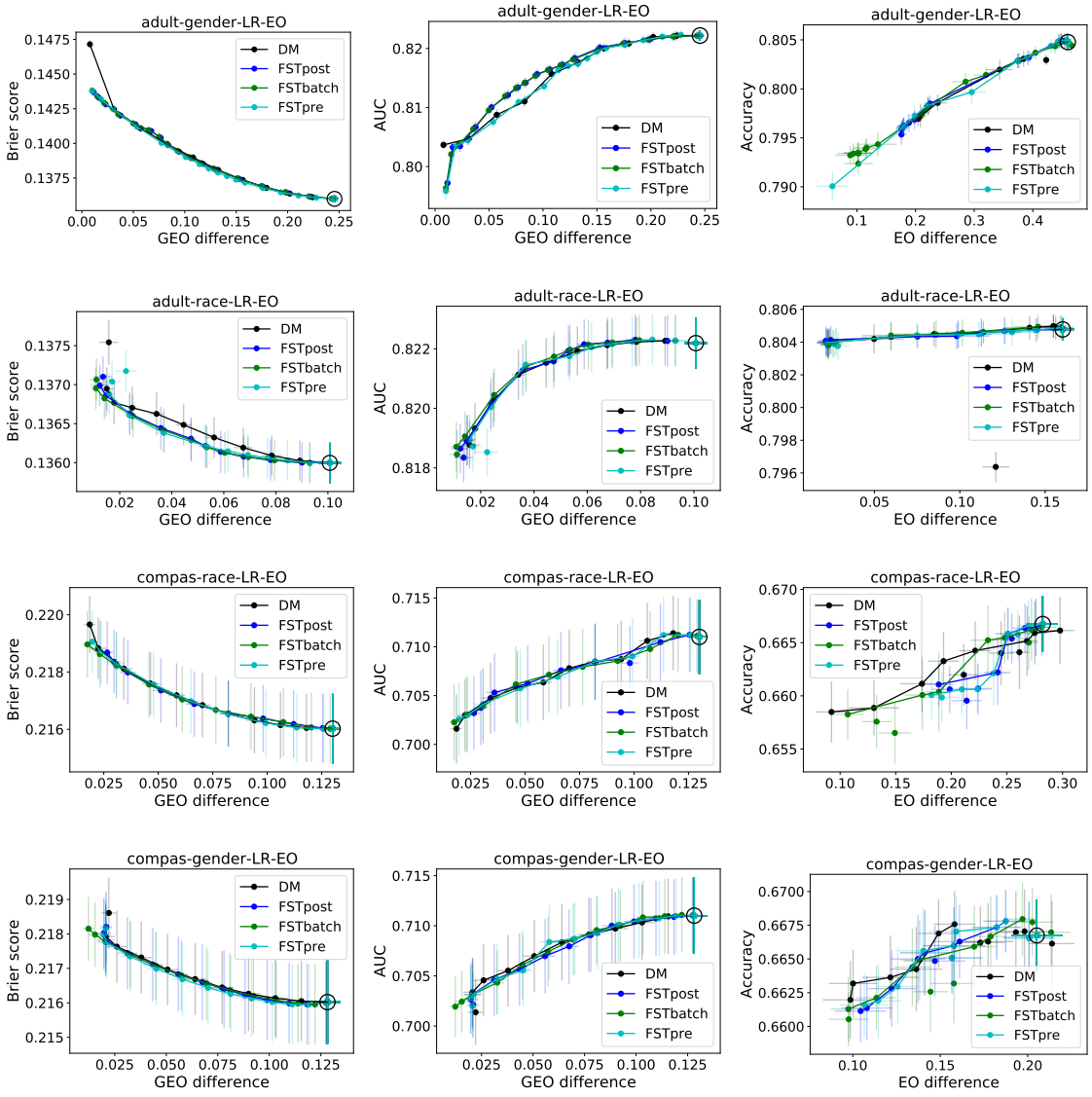



Figure 15: Trade-offs between equalized odds and classification performance measures for the Adult and COMPAS data sets with a reduced set of features.

the Adult and COMPAS data sets. On COMPAS, all methods are remarakably similar while on Adult, DM might be slightly worse. For example on Adult-gender (first row), DM does not reduce the EO difference below 0.2 (right panel). We reiterate however our lack of success with DM on full-dimensional data sets.

We also compare FST to the Fair Empirical Risk Minimization (FERM) approach of Donini et al. (2018). We use the code[3] provided by the authors. FERM, although a gen-

3. `https://github.com/jmikko/fair_ERM`

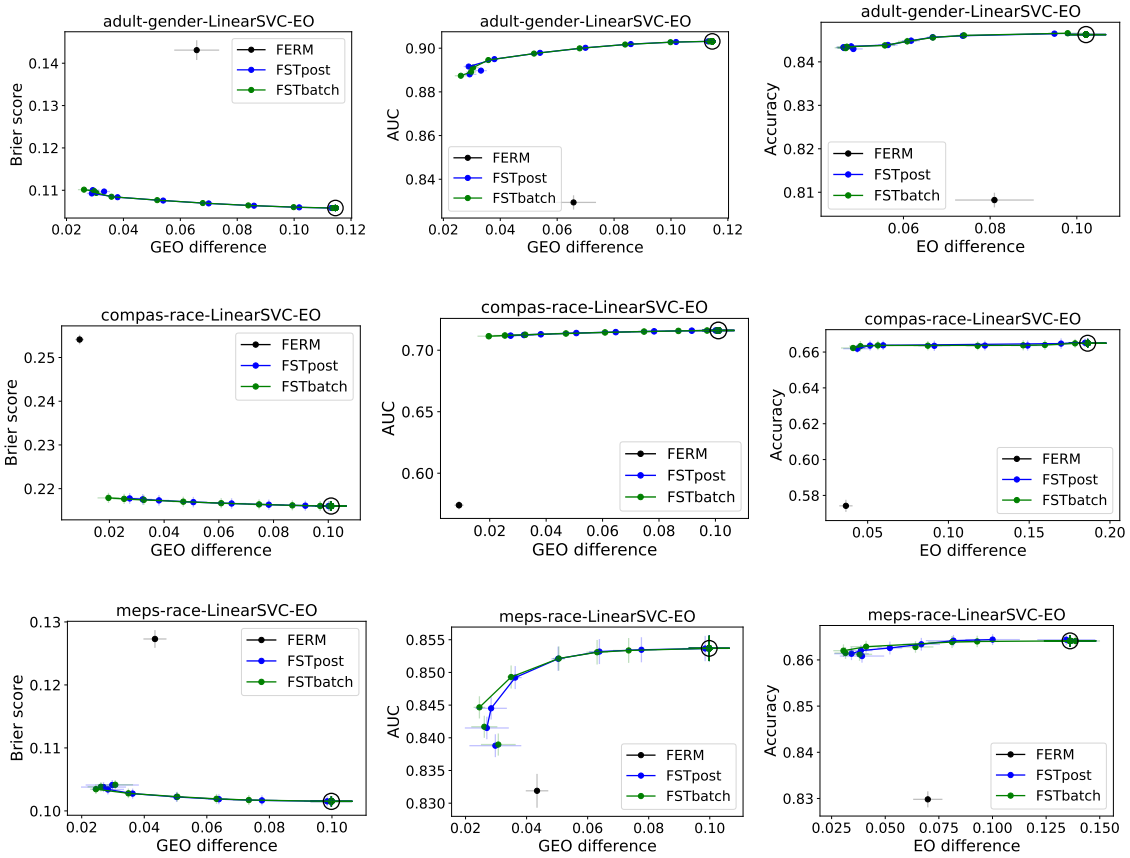



Figure 16: Trade-offs between fairness and classification performance measures for FERM (Donini et al., 2018) and our proposed FST approaches.

eral principle, has been specified only for binary classification problems with hinge loss as the loss function, and equal opportunity as the fairness constraint in Donini et al. (2018). The code provided by the authors implements linear and kernel support vector classifiers (SVC) with an equal opportunity constraint between two protected groups. During our experimentation, we observed that kernel SVC formulations of FERM were computationally impractical for the data sets we used (Adult, COMPAS, and MEPS). For example, experiments with the Adult data set using the RBF kernel SVC formulation did not finish even after waiting for 24 hours, whereas the linear formulation took only minutes to complete.[4] We suspect that this is because the kernel SVC formulation is implemented using a generic convex optimization solver[5] that does not incorporate any techniques for speedup specific to the problem. Hence we report results only for the linear SVC formulation. We also note that we use equalized odds as the fairness constraint in FST, which is stricter than the equal opportunity constraint used by FERM. These comparisons are illustrated in Figure

4. Experiments were performed on a machine running Ubuntu OS with 32 cores, and 64 GB RAM.
5. `http://cvxopt.org/`

16. Clearly, our FST methods that post-process probability outputs from linear SVC (FST-post, FSTbatch) outperform FERM substantially. We note however that the pre-processing variant of FST (FSTpre), which trains a second linear SVC model using sample weights described in Section 4.4, did not provide acceptable results. One possibility is that these sample weights, which are based on conditional probabilities, do not work well with the SVC problem formulation which is non-probabilistic. Nevertheless, in general we see that among all the four in-processing approaches we compared, only the reductions approach (Agarwal et al., 2018) has performance competitive to ours.

## References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning (ICML)*, pages 60–69, July 2018. URL `http://proceedings.mlr.press/v80/agarwal18a.html`.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, USA, 2nd edition, 1999.

Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA, USA, 2nd edition, 1997.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, pages 1–5, August 2017.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, September 2010. doi: 10.1007/s10618-010-0190-x. URL `https://doi.org/10.1007/s10618-010-0190-x`.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Optimized pre-processing for discrimination prevention. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3992–4001, December 2017. URL http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf.

L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 319–328, January 2019. doi: 10.1145/3287560.3287586. URL http://doi.acm.org/10.1145/3287560.3287586.

Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence (AAAI)*, January 2019.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017. doi: 10.1089/big.2016.0047.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 797–806, August 2017. doi: 10.1145/3097983.3098095. URL http://doi.acm.org/10.1145/3097983.3098095.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *AAAI/ACM Conference on Artifical Intelligence, Ethics, and Society (AIES)*, pages 1–8, January 2019.

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

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 2791–2801, December 2018. URL http://papers.nips.cc/paper/7544-empirical-risk-minimization-under-fairness-constraints.pdf.

John Duchi. Probability bounds. https://stanford.edu/~jduchi/projects/probability_bounds.pdf.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226. ACM, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency (FAccT)*, pages 119–133, February 2018. URL `http://proceedings.mlr.press/v81/dwork18a.html`.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*, pages 1–14, May 2016.

Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, August 2015. doi: 10.1145/2783258.2783311. URL `http://doi.acm.org/10.1145/2783258.2783311`.

Benjamin Fish, Jeremy Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *SIAM International Conference on Data Mining (SDM)*, pages 144–152, May 2016.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 329–338. ACM, 2019.

Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, July 2013. doi: 10.1109/TKDE.2012.72.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 3323–3331, December 2016. URL `http://dl.acm.org/citation.cfm?id=3157382.3157469`.

Hoda Heidari, Claudio Ferrari, Krishna Gummadi, and Andreas Krause. Fairness behind a veil of ignorance: A welfare analysis for automated decision making. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 1265–1276, December 2018. URL `http://papers.nips.cc/paper/7402-fairness-behind-a-veil-of-ignorance-a-welfare-analysis-for-automated-decision-making.pdf`.

José Hernández-Orallo, Peter Flach, and Cèsar Ferri. A unified view of performance metrics: Translating threshold choice into expected classification loss. *Journal of Machine Learning Research*, 13:2813–2869, October 2012.

Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, June 2020.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1–10, July 2019.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012. doi: 10.1007/s10115-011-0463-8. URL https://doi.org/10.1007/s10115-011-0463-8.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *IEEE International Conference on Data Mining (ICDM)*, pages 924–929, Dec 2012. doi: 10.1109/ICDM.2012.45.

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, June 2013. doi: 10.1007/s10115-012-0584-8. URL https://doi.org/10.1007/s10115-012-0584-8.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 35–50, September 2012.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*, pages 2569–2577, 2018.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 656–666, December 2017. URL http://dl.acm.org/citation.cfm?id=3294771.3294834.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, pages 43:1–43:23, 2017.

Emmanouil Krasanakis, Eleftherios Spyromitros-Xioufis, Symeon Papadopoulos, and Yiannis Kompatsiaris. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *International World Wide Web Conference (WWW)*, pages 853–862, 2018. doi: 10.1145/3178876.3186133.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 4066–4076, December 2017. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer Verlag, Berlin, Heidelberg, 1991.

Percy Liang. CS229T/STAT231 Statistical Learning Theory lecture notes, April 2016. https://web.stanford.edu/class/cs229t/notes.pdf.

Gregory YH Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry JGM Crijns. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial

fibrillation using a novel risk factor-based approach: the Euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272, 2010.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair encoder. In *International Conference on Learning Representations (ICLR)*, pages 1–11, May 2016.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*, pages 3384–3393, July 2018. URL `http://proceedings.mlr.press/v80/madras18a.html`.

Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency (FAccT)*, pages 107–118, February 2018. URL `http://proceedings.mlr.press/v81/menon18a.html`.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 1931–1940, February 2018. URL `https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16683/15898`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. URL `http://jmlr.org/papers/v12/pedregosa11a.html`.

Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *ACM Symposium on Applied Computing*, pages 126–131. ACM, 2012.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 5680–5689, December 2017. URL `http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf`.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *ACM SIGMOD/PODS International Conference on Management of Data (SIGMOD)*, pages 793—-810, June 2019. doi: 10.1145/3299869.3319901.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P. Calmon. Optimized score transformation for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, June 2020.

Blake Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory (COLT)*, pages 1920–1953, July 2017. URL http://proceedings.mlr.press/v65/woodworth17a.html.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 585–596, December 2017. URL http://papers.nips.cc/paper/6661-controllable-invariance-through-adversarial-feature-learning.pdf.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (BigData)*, pages 570–575, December 2018. doi: 10.1109/BigData.2018.8622525.

Forest Yang, Moustapha Cisse, and Oluwasanmi Koyejo. Fairness with overlapping groups. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2020.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International World Wide Web Conference (WWW)*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Rodriguez, Krishna Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. In *Conference on Neural Information Processing Systems (NeurIPS)*, pages 229–239, December 2017b. URL http://papers.nips.cc/paper/6627-from-parity-to-preference-based-notions-of-fairness-in-classification.pdf.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 962–970, April 2017c. URL http://proceedings.mlr.press/v54/zafar17a.html.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pages 325–333, Atlanta, Georgia, USA, June 2013. URL http://proceedings.mlr.press/v28/zemel13.html.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on Artifical Intelligence, Ethics, and Society (AIES)*, pages 335–340, February 2018. doi: 10.1145/3278721.3278779. URL http://doi.acm.org/10.1145/3278721.3278779.

Zirun Zhao, Anne Chen, Wei Hou, James M Graham, Haifang Li, Paul S Richman, Henry C Thode, Adam J Singer, and Tim Q Duong. Prediction model and risk scores of ICU admission and mortality in COVID-19. *PLOS ONE*, 15(7):e0236618, 2020.