Surrogate Regret Bounds for Equalized Odds

William Brown

Abstract:

We consider the problem of fairness in machine learning from the perspective of statistical consistency. We focus on the equalized odds criterion, which mandates equality of TPR and FPR for protected groups. The set of feasible equalized odds predictors is defined by the intersection of the AUC regions for all groups. For the simple post-processing procedure from Hardt et al. [6], we obtain a surrogate regret bound for equalized odds classifiers via strongly proper loss functions. Our proof makes use of existing results relating classification regret to AUC regret, as well as surrogate regret bounds for AUC via proper losses.

1 Introduction

The emergent field of fairness in machine learning has arisen out of concerns that trained models may result in discriminatory behavior when deployed in practice. The use of machine learning for tasks with legal protections against discriminatory behavior (such as hiring, credit allocation, and criminal sentencing) is relatively commonplace, and has resulted in observed disparate impact to various social groups. Notably, a study conducted by ProPublica found that the widely used COMPAS software for recidivism prediction results in a disproportionately high false positive rate for labeling black inmates as being likely to reoffend [3]. This type of disparate error can occur even when the training algorithm does not have access to the sensitive features (such as race, gender, income, etc.); many other features may be sufficiently correlated with the protected attributes such that the algorithm may effectively (and unintentionally) predict the protected attribute and make use of it in classification [4]. These observations have prompted a great deal of debate about fairness in machine learning. In many such cases, we would desire learning algorithms which provide some guarantee that its classification decisions are somehow being made fairly. It is not at all clear how we ought to define fairness in classification settings, or whether it is even possible for an algorithm to be completely fair. Notions have been defined which focus on individual fairness, where a goal is to treat similar individuals similarly, as well as group fairness, where the aim is to prevent discrimination with respect to a sensitive attribute, such as race or gender [4]. For group fairness criteria, it is often impossible to satisfy several notions of fairness simultaneously, except in extreme edge cases (such as having a perfect classifier)[8]. However, many of the most prominent fairness criteria are statistical in nature. Groups of the population are defined in terms of some protected attribute, and the criteria state that some statistical function of the classifier for each group, such as false positive rate or positive predictive value, should be (approximately) equalized across all relevant groups.

Within the field of statistical learning theory, much attention has been given over the last decade and more to notions of statistical consistency in learning, where the aim is to derive bounds on the regret of a hypothesis obtained by minimizing some loss function. Any distribution will have some non-negative Bayes error for a given loss function, indicating the minimum possible error
under the loss function achievable by any hypothesis. We typically would like to get as close to this error as possible, absent computational restrictions. It is often the case that the true relevant loss function for some task (such as 0-1 loss for binary classification) is computationally intractable to minimize directly [5]. Instead, it is common to make use of a surrogate loss function, for which there exist efficient algorithms for empirical risk minimization. For some families of surrogate loss functions, regret transfer bounds can be derived between the surrogate and target loss functions, which show that diminishing regret for the surrogate loss implies diminishing regret for the target loss, as the training sample size goes to infinity. For example, the family of $\lambda$-strongly proper composite losses admit regret transfer bounds for the 0-1 loss, via a link function for interpolating between the surrogate and target prediction spaces. An algorithm making use of such a function thus obtains a regret guarantee with respect to the Bayes error as the sample size increases. These types of guarantees are important for characterizing the robustness of learning procedures.

Our aim in this work is to elucidate connections between these two threads of research. Established metrics for fairness in learning involve approximately equalizing the label-restricted error rates for different groups of a population. We will focus primarily on the equalized odds condition from Hardt et al. [6], which demands simultaneous equality of true positive rates and false positive rates for each protected group. This criterion has the appealing semantic property of addressing both Type 1 and Type 2 errors, lending an intuitive notion of fairness to a variety of scenarios where one type of error may be seen as more costly than the other. It also has a straightforward technical interpretation which makes it relatively simple to satisfy via post-processing (albeit possibly with non-trivial loss in accuracy) if the hypothesis is allowed to differ for each protected group. Using different classifiers for each group is sometimes unavoidable; Dwork et al. [4] show that making use of group labels in classification can be necessary for achieving fairness, as different relationships between features and labels for each group can preclude the possibility of an effective one-size-fits-all classifier.

Suppose we have some trained predictor $f$ where $f(X)$ increases monotonically with the likelihood that the true label $Y(X) = 1$, and we make our predictions using a threshold $t$, where $\hat{Y}(X) = 1(f(Y) > t)$. Varying the value of $t$ over the range of $f$ when limiting the distribution to samples from a certain group traces out the ROC curve achievable for that group by $f$. Each point on the curve results in a distinct false positive and true positive rate when the corresponding threshold is used. As ROC curves are concave, any point within the region between the curve and the (0,0), (1,1) line is achievable by randomizing between thresholds. Given ROC curves for every group, minimizing loss while satisfying equalized odds is equivalent to finding an optimal point in the intersection of the area under the ROC curve for each group and using (possibly randomized) thresholds for each group corresponding to that point. This post-processing technique for real-valued predictors, presented by Hardt et al. [6], is straightforward to implement and requires very few assumptions about the original predictor.

Binary classification with a fairness requirement is effectively a constrained variant of the standard binary classification optimization problem with a discrete loss. Statistically consistent learning procedures are well-studied for the unconstrained task, but much more remains unknown for the case when the trained classifier must satisfy a fairness metric. From the perspective of statistical consistency, a natural goal would be to derive regret bounds for this approach with respect to the optimal fair classifier for some class of predictors. Our main result is a surrogate regret bound for predictors satisfying equalized odds via strongly proper losses. Our technique makes use of existing
regret bounds for the AUC performance metric as an intermediary between a surrogate loss and the
target discrete loss, and we show that these bounds can be applied to the feasible region of equalized
odds predictors. This result serves as a partial answer to a call from Menon and Williamson [9] to
establish consistency guarantees for thresholding approaches to fair classification. Our results hold
for cost-sensitive classification and for multi-category sensitive features, although the strength of
our regret bound will depend on the number of protected groups.

2 Preliminaries

Receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC) are
central to our study; the ROC curves for each group are the constraints for our optimization
problem, and existing regret bounds for AUC give us an approach towards a consistent optimization
procedure. We recall several important details about equalized odds, as well as the consistency
guarantees related to AUC which will be useful for developing our results.

2.1 ROC Curves

ROC curves illustrate the tradeoffs between label-restricted error rates as a classifier’s sensitivity is
varied. If a classifier outputs a real-valued score, varying the threshold for a positive classification
results in different true positive and false positive rates, tracing out the ROC curve. AUC is an
important measure of performance for learning tasks. If a distribution has $D$ non-zero Bayes error
or risk, the optimal AUC ($\text{AUC}_D^*$) defines the limits of the error tradeoffs which are achievable,
given infinite data and computation. Notions of statistical consistency for various convex loss
functions provide useful tools for minimizing generalization error when data is limited and the true
optimization problem is non-convex. Agarwal [1] shows the AUC-consistency of strongly proper
loss functions and provides a regret bound, allowing us to obtain quantitative guarantees in terms
of the chosen surrogate loss function. Balcan et al. show an equivalence between optimizing AUC
and accuracy in bipartite ranking [2], and Narasimhan and Agarwal [11] give regret bounds for
discrete cost-sensitive loss in terms of the loss for the bipartite ranking problem. While regret
bounds for discrete loss can be shown directly in terms of a surrogate loss [12], a similar bound
can also be obtained using the AUC regret bound as an intermediary. Our work follows the latter
approach.

2.2 Equalized Odds

Hardt et al. introduce equalized odds in [6] as a fairness criterion for binary classification. In this
setting we are tasked with predicting $Y$ from $(X, A)$ for a joint distribution $D$ over $(X, Y, A)$, with
features $X$, true labels $Y \in \{0, 1\}$, and a protected attribute $A$. We refer to the marginal distribu-
tion for a given protected attribute as a protected group. Informally, equalized odds demands that
both the True Positive Rate (TPR) and False Positive Rate (FPR) for all protected groups must
be equalized.

**Definition 1** (Equalized Odds for Binary Predictors [6]). A predictor $\hat{Y}$ satisfies equalized odds
(EO) with respect to protected attribute $A$ and outcome $Y$, if $\hat{Y}$ and $A$ are independent conditional
on $Y$. Equivalently,
\[
\Pr[\hat{Y} = 1 \mid A = 0, Y = y] = \Pr[\hat{Y} = 1 \mid A = 1, Y = y], \quad y \in \{0, 1\}
\]

Hardt et al. focus on the case where the protected attribute is binary. For more than two protected groups, we can also define equalized odds for all $a$ and $a'$ in $A$:
\[
\Pr[\hat{Y} = 1 \mid A = a, Y = y] = \Pr[\hat{Y} = 1 \mid A = a', Y = y], \quad y \in \{0, 1\}, a, a' \in A
\]

### 2.2.1 Feasible Predictors

Hardt et al. describe a method for attaining a predictor which satisfies equalized odds given a real-valued score $R \in [0, 1]$, such as a class probability estimator. This technique makes use of the ROC curves yielded by $R$ for each protected group, characterizing the tradeoffs between TPR and FPR.

**Definition 2** (A-Conditional ROC Curves [6]). The $A$-conditional ROC curve for a predictor $R$ is a function $C_a(t) : [0, 1] \rightarrow [0, 1]^2$ parameterized by a threshold $t$, where:
\[
C_a(t) = (\Pr[R > t \mid A = a, Y = 0], \Pr[R > t \mid A = a, Y = 1])
\]

This curve defines the feasible region of classifiers achievable for a group $a$ with $R$ in TPR/FPR space. Such a classifier is given by $\hat{Y} = \mathbb{I}(R > t)$ for some $t \in [0, 1]$.

**Lemma 1** (Feasible Regions [6]). For every group $a$, the convex hull $H_a$ of the image of the $a$-conditional ROC curve
\[
H_a = \text{convhull}(C_a(t) : t \in [0, 1])
\]
is equivalent to the set of all non-trivial TPR/FPR pairs which are feasible for $a$ with $R$. Any point within the convex hull is achievable by randomizing between thresholds. All points below the line between $(0, 0)$ and $(1, 1)$ are omitted as they represent behavior worse than an entirely random classifier which is independent of $X$.

The feasible region for equalized odds classifiers over all groups is given by the intersection of all feasible regions, i.e. $\bigcap_a H_a$. Hardt et al. note that optimizing over this region is feasible via ternary search. They also show that this approach is essentially optimal, given an optimal scoring function.

**Theorem 1** ([6]). For any source distribution over $(Y, X, A)$ with Bayes optimal regressor $R(X, A)$, and any loss function $\ell$, there exists a predictor $Y^*(R, A)$ such that:

- $Y^*$ is an optimal predictor satisfying equalized odds. That is, $\mathbb{E}[\ell(Y^*, Y)] \leq \mathbb{E}[\ell(\hat{Y}(X, A), Y)]$ for any predictor $\hat{Y}(X, A)$ which satisfies equalized odds.
- $Y^*$ is derived from $(R, A)$, that is, it is solely a function of $(R, A)$ and conditionally independent of $X$ given $(R, A)$.

Our main result shows the statistical consistency of this thresholding approach with the Bayes-optimal equalized odds predictor when $R$ is trained by optimizing an appropriate surrogate loss function.
2.3 Transfer Bounds Between Classification and Ranking

2.3.1 Cost-Sensitive Binary Classification

The goal of standard binary classification is to select a hypothesis which minimizes cost-sensitive prediction error. We provide the corresponding definitions of error and regret.

**Definition 3** (Cost-Sensitive Error [11]). The cost-sensitive 0-1 error with cost parameter $c$ for a hypothesis $h : X \rightarrow Y$ and a distribution $D$ over $(X,Y)$ is given by

$$er_{D}^{0-1,c}[h] = \mathbb{E}_{(x,y) \sim D}[(1 - c)\mathbb{I}(y = 1, h(x) = -1) + c\mathbb{I}(y = -1, h(x) = 1)]$$

where $\mathbb{I}(\cdot)$ is the boolean identity function taking value 1 if its argument is true and 0 if false.

**Definition 4** (Cost-Sensitive Regret [11]). The cost-sensitive regret of a hypothesis $h$ is given by the difference between its error and the error of the best classifier for the distribution

$$\text{regret}_{D}^{0-1,c}[h] = er_{D}^{0-1,c}[h] - er_{D}^{0-1,c,*}$$

where $er_{D}^{0-1,c,*} = \inf_{h \colon X \rightarrow Y} er_{D}^{0-1,c}[h]$.

2.3.2 Bipartite Ranking

The goal in bipartite ranking is to learn a ranking function $f : X \rightarrow \mathbb{R}$ where the score for an instance increases monotonically with the likelihood that the true label is positive. The optimization objective is to minimize the likelihood that a pair of instances are ranked incorrectly. We can again define notions of error and regret for the ranking problem.

**Definition 5** (Ranking Error [11]). The ranking error of a function $f$ and a distribution $D$ is given by

$$er_{D}^{\text{rank}}[f] = \mathbb{E}_{(x,y),(x',y') \sim D}[(\mathbb{I}(y - y')(f(x) - f(x')) < 0) + \frac{1}{2}\mathbb{I}(f(x) = f(x') \mid y \neq y')]$$

**Definition 6** (Ranking Regret [11]). The ranking regret of a function $f$ and a distribution $D$ is given by

$$\text{regret}_{D}^{\text{rank}}[h] = er_{D}^{\text{rank}}[h] - er_{D}^{\text{rank,*}}$$

where $er_{D}^{\text{rank,*}} = \inf_{f \colon X \rightarrow \mathbb{R}} er_{D}^{\text{rank}}[h]$.

2.3.3 Regret Transfer Bounds

Narasimhan and Agarwal show regret transfer bounds between these two problems, where low ranking regret implies low classification regret. Their result relies on a weak distributional assumption (Assumption A henceforth), namely that the induced distribution of scores from $(D, f)$ is either discrete, continuous, or mixed with finitely many point masses.
Theorem 2 (Regret Transfer Bound from Ranking to Classification [11]). For \( D, f, \) and cost parameter \( c \), let \( t^* \) denote the optimal threshold for \( f \), where

\[
t^* = \arg\min_{t \in \mathbb{R}} \text{er}_D^{0-1,c}[\text{sign}(f - t)]
\]

When Assumption A holds, the classification regret can be bounded by

\[
\text{regret}_D^{0-1,c}[\text{sign}(f - t^*)] \leq \sqrt{2p(1-p)\text{regret}_{\text{rank}}^D[f]}
\]

Furthermore, when \( t^* \) is not directly accessible and must be approximated via optimization over a sample of \( n \) points from \( D \), the regret bound worsens only by an additive term which is a function of a confidence parameter \( \delta \) and decays at rate \( \tilde{O}(\frac{1}{\sqrt{n}}) \).

2.4 Equivalence of AUC and Ranking Optimization

As noted by Balcan et al. [2] and others, the problem of minimizing loss in ranking is equivalent to maximizing AUC for a scoring function. More specifically, \( \text{er}_D^{\text{rank},*} = 1 - \text{AUC}_D^* \) and \( \text{er}_D^{\text{rank},f} = 1 - \text{AUC}_D[f] \). We can define a notion of regret for AUC and directly apply Theorem 2 to the AUC maximization problem.

Definition 7 (AUC Regret). The AUC regret of a scoring function \( f : \mathcal{X} \to \mathbb{R} \) is

\[
\text{regret}_D^{\text{AUC}}[f] = \text{AUC}_D^* - \text{AUC}_D[f]
\]

Theorem 3 (Regret Transfer Bound from AUC to Classification [11]). For \( D, f, \) and cost parameter \( c \), and optimal threshold \( t^* \), when Assumption A holds, the classification regret can be bounded by

\[
\text{regret}_D^{0-1,c}[\text{sign}(f - t^*)] \leq \sqrt{2p(1-p)\text{regret}_D^{\text{AUC}}[f]}
\]

Note that we can also define AUC using the machinery from Hardt et al.:

\[
\text{AUC}_D[f] = H_f + \frac{1}{2}
\]

where \( H_f \) is the group-agnostic feasible region induced by \( f \) and the added \( \frac{1}{2} \) term corresponds to the omitted region below the \(((0,0), (1,1))\) line.

2.5 Surrogate Regret Bounds for AUC via Strongly Proper Losses

Agarwal [1] shows a regret bound for AUC via strongly proper composite loss functions. The family of strongly proper composite losses includes many common loss functions such as the exponential, logistic, and squared loss. For further technical background, we direct the reader to the paper from Agarwal [1] which introduced the notion of strongly proper losses. Our regret bounds will be derived in terms of this class of loss functions.
Theorem 4 (AUC Regret Bound). Let \( \hat{Y} \subseteq \bar{\mathbb{R}} \), let \( p = \Pr[Y = 1] \), and let \( \lambda > 0 \). Let \( \ell : \{\pm 1\} \times \hat{Y} \to \bar{\mathbb{R}}_+ \) be a \( \lambda \)-strongly proper composite loss. Then for any \( f : \mathcal{X} \to \hat{Y} \),

\[
\text{regret}^\text{AUC}_D[f] \leq \frac{\sqrt{2}}{p(1 - p)\sqrt{\lambda}} \sqrt{\text{regret}_D^\ell[f]}
\]

Theorems 3 and 4 will serve as important tools in establishing our regret bounds for equalized odds.

3 Regret Bounds for Equalized Odds

We are now ready to prove our main results, which will make use of the strongly proper composite losses from Agarwal [1]. We first outline the learning procedure for which we will derive our regret bounds.

3.1 Learning Procedure

Given access to samples \((X, Y, A)\) from a distribution \(D\):

- Partition the samples by protected group into sets \(S_a\) for all \(a\).
- Select a learning algorithm which corresponds to minimizing \(\lambda\)-strongly proper composite loss, such as AdaBoost for the exponential loss or logistic regression for the logistic loss.
- Train a scoring function \(f_a\) for each group using the samples in \(S_a\), resulting in a feasible region \(H_a\).
- Compute the intersection of all feasible regions \(\bigcap_a H_a\) and optimize thresholds for cost-sensitive loss via ternary search, as described in [6].

While we are training on disjoint sample sets for each classifier, it is straightforward to see that the optimal AUC for each group, and thus the optimal feasible region for each group, is the same regardless whether samples from other groups can be used for training. Recall the Bayes AUC is a supremum over all scoring functions; this includes scoring functions which are allowed to read a transcript which exactly characterizes the full joint distribution \(D\) during training. This implies that while transfer learning between groups may provide practical improvements on finite samples, it is not needed for statistical consistency.

3.2 AUC Intersection Regret

We define notions of optimality and regret for the feasible region of equalized odds predictors in terms of the intersection of AUC (IntAUC) for all groups.

**Definition 8** (Bayes Optimal IntAUC). The Bayes optimal AUC intersection for all groups is given by

\[
\text{IntAUC}^*_D = |\bigcap_a H_a^*| + \frac{1}{2}
\]

7
where $H_a^*$ is the convex hull above the $((0,0),(1,1))$ line corresponding to the Bayes optimal scoring function which produces AUC$_D^*$. Here we slightly abuse notation and let $|\bigcap_a H_a^*|$ denote the Euclidean area for the convex hull $\bigcap_a H_a^*$. Convexity of $\bigcap_a H_a^*$ follows from the fact that the intersection of convex regions is convex.

**Definition 9 (IntAUC Regret).** For a set of scoring functions $F = \{f_a : a \in A\}$, regret for the AUC intersection for all groups is given by

$$
\text{regret}_{D}^{\text{IntAUC}}[F] = \text{IntAUC}_{D}^* - \text{IntAUC}_{D}[F] = \text{IntAUC}_{D}^* - (|\bigcap_a H_{a,f_a}| + \frac{1}{2})
$$

We can additionally show that $\text{regret}_{D}^{\text{IntAUC}}[F]$ is bounded by the individual AUC regrets for each group.

**Lemma 2.** For a set of scoring functions $F$ and distribution $D$,

$$
\text{regret}_{D}^{\text{IntAUC}}[F] \leq \sum_a \text{regret}_{D}^{\text{AUC}}[f_a]
$$

**Proof.**

$$
\text{regret}_{D}^{\text{IntAUC}}[F] = |\bigcap_a H_a^*| - |\bigcap_a H_{a,f_a}|
\leq |\bigcup_a (H_a^* \setminus H_{a,f_a})| (H_{a,f_a} \subseteq H_a^*)
\leq \sum_a |(H_a^* \setminus H_{a,f_a})| \quad (\text{union bound})
= \sum_a \text{regret}_{D}^{\text{AUC}}[f_a]
$$

The last line follows from the formulation of AUC in terms of the feasible reason for each group as discussed in Section 2.4.

With this, we can give a regret bound for IntAUC in terms of the regret for a proper loss, using Theorem 4.

**Theorem 5.** For a distribution $D$ over $(X,Y,A)$, let $p_a = \text{Pr}[Y = 1|A = a]$. For a set of scoring functions $F$ and $\lambda$-strongly proper composite loss $\ell$,

$$
\text{regret}_{D}^{\text{IntAUC}}[F] \leq \sum_a \frac{\sqrt{2}}{p_a(1 - p_a)\sqrt{\lambda}} \sqrt{\text{regret}_{D}[f_a]}
$$

**Proof.** This follows directly from applying Lemma 2 to Theorem 4.

8
3.3 Geometric Equivalence of IntAUC and AUC

Given a regret bound for IntAUC, our next step will be to derive classification regret bounds in terms of proper loss regret. The regret bound in Theorem 3 was derived using the machinery of bipartite ranking analysis. Before applying a version of this bound, we must justify that it also holds for IntAUC. We do so with a geometric argument; the bound in Theorem 3 is distribution-agnostic modulo Assumption A, and so by showing that any curve defining an intersection of feasible regions also defines the ROC curve for some distribution, we can interpret Theorem 3 as a geometric statement about concave functions.

**Lemma 3.** For any concave increasing function \( r : [0, 1] \to [0, 1] \) where \( r(0) = 0 \) and \( r(1) = 1 \), there is a distribution \( D \) for which \( r \) is equivalent to the optimal ROC curve \( C_D(t) \).

**Proof.** Let \( D \) be a distribution over \((X, Y)\) where \( X \in [0, 1] \) and \( Y \in \{0, 1\} \). For all \( x \) let \( \Pr[Y = 1] = x \), and let \( \Pr[X = x] = z(x) \), appropriately normalized such that \( \int_0^1 z(x)dx = 1 \). We will soon define \( z(x) \) more precisely. As the value \( X \) is equivalent to the optimal class probability estimator, the Bayes-optimal scoring function will simply be \( f(x) = x \). The true positive rate for \( f \) and a threshold \( t \) can be given as:

\[
TPR(t) = \frac{\int_t^1 xz(x)dx}{\int_0^1 xz(x)dx}
\]

The false positive rate for \( f \) and a threshold \( t \) can be given as:

\[
FPR(t) = \frac{\int_0^t (1 - x)z(x)dx}{\int_0^1 (1 - x)z(x)dx}
\]

\( FPR(t) \) and \( TPR(t) \) are both decreasing in \( t \), and are fully defined by \( z(x) \). Let \( z(x) \) be the probability density function such that \( r(TPR(t)) = FPR(t) \); we omit the full derivation of \( z(x) \) for brevity. This gives us our desired distribution.

Further, the cost-sensitive loss incurred by the optimal classifier for a given AUC region is symmetric to the loss incurred by the optimal equalized odds classifier within an identical AUC, given equal base rates and cost parameters.

**Theorem 6** (Regret Transfer Bound from AUC to Classification). For \( D \), cost parameter \( c \), a set of scoring functions \( F \), and the optimal set of thresholds \( T^* = \{t^*_a : a \in A\} \), when Assumption A holds for all \( f \in F \), the equalized odds classification regret can be bounded by

\[
\text{regret}^{0-1,EO,c}_D[\text{sign}(f - t^*)] \leq \sqrt{2p(1 - p)\text{regret}^{\text{IntAUC}}_D[f]}
\]

**Proof.** This follows from translating Theorem 3 to apply to intersections of AUC regions. From Lemma 2, the bound in Theorem 3 holds for any concave curve, and the intersection of AUC regions will be defined by some concave curve with endpoints at \((0, 0)\) and \((1, 1)\).
3.4 Equalized Odds Classification Regret Bounds

Our main result is a surrogate regret bound via strongly proper losses for classifiers satisfying equalized odds.

**Theorem 7** (Regret Bounds for EO Classification via Surrogate Losses). For \( D, \) cost parameter \( c, \) a set of scoring functions \( F, \) \( \lambda \)-strongly proper composite loss \( \ell, \) and the optimal set of thresholds \( T^* = \{ t^*_a : a \in A \}, \) when Assumption A holds for all \( f \in F, \) the equalized odds classification regret can be bounded by

\[
\text{regret}^{0-1, \text{EO},c}_D[\text{sign}(f - t^*)] \leq \sqrt{2p(1-p) \sum_a \frac{\sqrt{2}}{p_a(1-p_a)\sqrt{\lambda}} \sqrt{\text{regret}_D[f_a]}}
\]

**Proof.** This follows from substituting the surrogate loss bound from Theorem 5 for the AUC intersection regret in Theorem 6. \qed

This proves the statistical consistency of our aforementioned learning procedure with the best equalized odds classifier for a distribution. As was the case for the regret bound between ranking and classification, an additional \( \tilde{O}(\frac{1}{\sqrt{n}}) \) additive regret term will be incurred when sampling \( n \) points to optimize within the IntAUC region. Sampling for threshold selection also introduces the possibility that equality of opportunity is only approximately satisfied. Yet, this optimization only requires estimates of the group-specific base rates as well as TPR and FPR for a set of thresholds. All of these values can be answered via statistical queries, and so we can import well-known results from e.g., Kearns et al. [7] to obtain consistency for these estimations as well.

4 Future Directions

While statistical consistency for equalized odds shows optimality for our procedure as our sample size grows to infinity, the problem of identifying the optimal equalized odds classifier over a finite sample was shown to be computationally intractable by Woodworth et al. in [13]. As such, it is unlikely that we can obtain rates of convergence to the best fair classifier globally with this approach. However, rates of convergence to the best fair classifier within some constrained domain may be possible; Mukherjee et al. [10] give a rate of convergence for AdaBoost in terms of the exponential loss for classifiers with bounded parameter vectors. An equivalent formulation of our regret bounds in terms of a constrained set of hypotheses may be plausible as a result.

While our regret bound is specific to equalized odds, Hardt et al. give a similar post-processing technique in [6] for satisfying simpler fairness metrics such as equality of TPR or FPR in isolation. It may be possible to extend our results directly to these metrics.

References

[1] Shivani Agarwal. Surrogate regret bounds for the area under the roc curve via strongly proper losses. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual


