Consider the Alternatives: Navigating Fairness-Accuracy Tradeoffs via Disqualification

Guy N. Rothblum*
Weizmann Institute of Science
rothblum@alum.mit.edu

Gal Yona[†]
Weizmann Institute of Science
gal.yona@weizmann.ac.il

October 5, 2021

Abstract

In many machine learning settings there is an inherent tension between fairness and accuracy desiderata. How should one proceed in light of such trade-offs? In this work we introduce and study γ -disqualification, a new framework for reasoning about fairness-accuracy tradeoffs w.r.t a benchmark class $\mathcal H$ in the context of supervised learning. Our requirement stipulates that a classifier should be disqualified if it is possible to improve its fairness by switching to another classifier from $\mathcal H$ without paying "too much" in accuracy. The notion of "too much" is quantified via a parameter γ that serves as a vehicle for specifying acceptable tradeoffs between accuracy and fairness, in a way that is independent from the specific metrics used to quantify fairness and accuracy in a given task. Towards this objective, we establish principled translations between units of accuracy and units of (un)fairness for different accuracy measures. We show γ -disqualification can be used to easily compare different learning strategies in terms of how they trade-off fairness and accuracy, and we give an efficient reduction from the problem of finding the optimal classifier that satisfies our requirement to the problem of approximating the Pareto frontier of $\mathcal H$.

1 Introduction

Underlying the study of *algorithmic fairness* [Dwork et al., 2012, Hardt et al., 2016, Kusner et al., 2017, Kearns et al., 2018, Hébert-Johnson et al., 2017] in the context of supervised learning is the fundamental tension between different fairness desiderata and some other desired property, such as accurate predictions [Kleinberg et al., 2016, Chouldechova, 2017, Chen et al., 2018, Pleiss et al.,

^{*}This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17), from the U.S.-Israel Binational Science Foundation (grant 2018102), and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. Part of this work was done while the author was visiting Microsoft Research.

[†]This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819702), from the Israel Science Foundation (grant number 5219/17) and from the Simons Foundation Collaboration on the Theory of Algorithmic Fairness. This research was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center, and by a research grant from the Estate of Tully and Michele Plesser.

2017]. This tension exists for a broad array of non-discrimination requirements. For example, demographic parity (see e.g. [Dwork et al., 2012]) requires that the predicted labels are independent of group membership – a requirement that stands in direct tension with accuracy when when the base rates (expected labels) between the groups differ. In other cases, the tension is more subtle but still exists. For example, the notion of equal opportunity [Hardt et al., 2016] requires that the error rates of the classifier (false positives and/or false negatives) be equal across groups. While a perfect classifier that never errs does satisfy this requirement, in settings where perfect predictions are impossible (e.g. due to missing data or inherent uncertainty), satisfying error rate parity can stand in direct conflict with maximizing accuracy. Importantly, in practice the tension between fairness and accuracy may (and does) arise even when the learner is well-intentioned, e.g. because of missing data or constraints on the complexity of the learned predictor (which are necessary to ensure generalization).

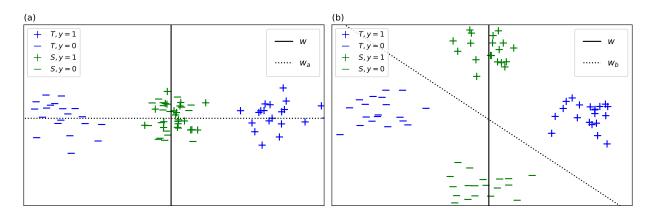


Figure 1: In both cases, the classifier $h_w(x) = \operatorname{sign}\langle w, x \rangle$ does not satisfy error rate parity, whereas the classifiers $h_{w_a}(x) = \operatorname{sign}\langle w_a, x \rangle$ (left) and $h_{w_b}(x) = \operatorname{sign}\langle w_b, x \rangle$ (right) do.

When fairness and accuracy are in conflict, as they often are, how can we reason about the possible trade-offs? If we require that error rates be balanced, this might preclude reasonable accuracy. We could opt, instead, to bound the error-rate disparity [Zafar et al., 2017], but what level of disparity would be appropriate? These considerations go well beyond the technical aspects of the problem at hand, and the answers must be informed by ethical, societal and legal considerations. In particular, any criterion that only considers a predictor in light of its error rate imbalance is inherently limited. Drawing inspiration from the legal literature, the following discussion of Disparate Impact doctrine from [Barocas and Selbst, 2016], p694, is instructive:

"Disparate impact is not concerned with the intent or motive for a policy; where it applies, the doctrine first asks whether there is a disparate impact on members of a protected class, then whether there is some business justification for that impact, and finally, whether there were less discriminatory means of achieving the same result"

Consider the tension between accuracy and fairness in light of this doctrine: we have disparate impact in the form of imbalanced error rates (or another measure of unfairness). Accuracy require-

 $^{^{1}}$ For example, suppose that the only data feature is group membership (in S or in T), that 80% of T have label 1 and 20% of S have label 1. For non-trivial accuracy, we need to predict a 1-label with higher probability for members of S than for members of T, but this will result in a disparity in false negatives and false positives between the two groups.

ments might provide a business justification for this imbalance. However, we must also consider the alternatives, and whether there are less discriminatory means of achieving the same result.

We demonstrate this point through a simple example. We consider a simple setup with two groups S and T towards which we want to guarantee fair treatment. Consider the two scenarios described in Figure 1: individuals from T are denoted in blue, and individuals from S are denoted in green. The learning task at hand is finding a low-error hyperplane. In both cases the data for T is perfectly separable, and the hyperplane w, which is tailored for T, has no false positives or false negatives on T. On the set S, however, the hyperplane w performs poorly (in both scenarios): half of its classifications will be false positives or false negatives. Given the disparity in error rates between S and T, should we rule out the predictor w for being discriminatory? Here, the two examples may diverge. In scenario (b) the hyperplane w is clearly and blatantly unfair: the hyperplane w_b , which takes S into account, has zero error-rate disparity, and is no less accurate than w (indeed, its loss is smaller). In scenario (a), on the other hand, the data of S is inseparable and there is no clear alternative to w. The only clear alternative with balanced error rates is w_a , which has huge error. While error rate disparity is an intuitively appealing measure, on its own it does not allow us to make distinctions between very different scenarios, such as (a) and (b).

Exact disqualification. The example discussed above suggests a natural first step towards formalizing the disparate impact doctrine: a classifier should be "disqualified" if there is an alternative classifier with similar accuracy that is more fair, as was the case for the hyperplane w in scenario (b) above. Importantly, this requires that we specify a set \mathcal{H} of alternatives to be considered. Most naturally, \mathcal{H} could be the hypothesis class used by a learning procedure. However, there is an important subtlety here: what do we mean by an alternative classifier being "more fair"? We might define the amount of unfairness in absolute value, e.g. the absolute value of the error disparity between the two groups, but this is problematic. Consider two equally accurate classifiers h_S and h_T , where h_g denotes a classifier that has an error disparity in favour of group $g \in \{S, T\}$. When measuring unfairness in absolute value, neither classifier disqualifies the other! But this is not what we want, as intuitively, neither classifier should be "kosher" to use: For example, from the perspective of S, it is unjustified to use h_T (which has a high error rate on S) when there is an equally-accurate classifier that S prefers. Similarly, T could make a similar argument against h_S . This discussion highlights that in order to prevent unfair treatment towards both groups, our notion of fairness must carefully take into account the *directionality* of the fairness violations.

Beyond Exact Disqualification. Exact disqualification gives us a framework to compare a given classifier to a set of available alternatives, but it does not fully resolve the question of navigating fairness and accuracy tradeoffs because it leaves many classifiers incomparable. For example, if relative to a classifier h, a classifier h' improves fairness by τ but hurts accuracy by τ' , neither

²While we focus on the two-group setup, our approach naturally lends itself to a "multi-group" extension [Hébert-Johnson et al., 2017, Kearns et al., 2018], where instead of only two groups we consider a much a larger collection of groups; see the discussion in Section 5.

 $^{^3}$ We emphasize that we do not assert the hyperplane w is a "fair" classifier in scenario (a): this is a task-specific and context-dependent question. Our main point is that studying the set of alternative classifiers is instructive, and points to the glaring unfairness of using w in scenario (b). Moreover, studying these alternatives can also be instructive in scenario (a): it may lead to the conclusion that we need to collect new data features or consider a richer hypothesis class.

classifier disqualifies the other. Choosing between such classifiers implicitly requires comparing a fairness gain of τ to an accuracy loss of τ' . This comparison is not straightforward, since these quantities are typically not measured in the same "units" - for example, accuracy is usually averaged over the entire population whereas fairness typically quantifies some across-group disparity. This difficulty is further exacerbated by the fact that both fairness and accuracy are in general two abstract concepts, that can be operationalized in different ways (e.g. by choosing different metrics based on the task in question).

We address these challenges by quantifying the relative importance of fairness vs. accuracy. In particular, our notion of γ -disqualification stipulates that a classifier should be disqualified if there is a fairer alternative that does not degrade accuracy "too much". This is more restrictive than exact disqualification: we enlarge the set of disqualifying alternatives, by including some that do degrade accuracy, so long as the accuracy degradation is not "too much" larger than the improvement in fairness. The notion of "too much" is quantified using a parameter $\gamma \geq 0$. Importantly, our objective is for γ to reflect the acceptable tradeoffs between fairness and accuracy in a way that is independent of the specific metrics chosen to operationalize these concepts. As discussed above, this requires specifying an appropriate normalization to first bring fairness and accuracy to the same "units", so the parameter γ can meaningfully specify acceptable accuracy-fairness tradeoffs. This is a crucial ingredient (and contribution) of our framework. The larger γ is, the more weight we give to any degradation in the classifier's accuracy, and the less restrictive the requirement becomes:

1 unit of accuracy $\equiv \gamma$ units of fairness 1 unit of fairness $\equiv 1/\gamma$ units of accuracy

In particular, setting $\gamma=\infty$ recovers exact disqualification, whereas $\gamma=0$ means that we require that the classifier be optimally fair within \mathcal{H} . For example, if the fairness notion we use is parity of error rates and the class \mathcal{H} includes a balanced classifier (e.g. the constant classifier), then setting $\gamma=0$ is equivalent to requiring error rate parity. The range of values $\gamma\in(0,\infty)$ specifies a range of acceptable fairness-accuracy trade-offs. We refer to the resulting requirement in short as γ -Fairness. Our framework can be instantiated with different loss functions, capturing different measures of accuracy, and their trade-offs with different measures of (un)fairness.

1.1 Related work

Other approaches to formalizing "disparate impact". We defer the full discussion of the different avenues by which the fairness community has proceeded in light of the basic tradeoffs between fairness and accuracy to Section 5, and here discuss how disqualification differs from perhaps the other natural approach to addressing the "business necessity" clause of the disparate impact doctrine, which is to maximize fairness subject to an accuracy constraint [Zafar et al., 2017]. They key difference is that this approach results in a model whose accuracy is not far from the accuracy of the optimal model, whereas such a property is not guaranteed in our framework, since it could be the case that all the non-disqualified models are significantly less accurate than the optimal one. However, we argue that this property is not necessarily desirable. In particular, the fact that the eventual model should be "not significantly less accurate than the optimal model" encodes a nor-

mative assumption, that may not be valid in all cases.⁴ Our framework handles such subtleties, because by definition, a model with significantly lower accuracy will only be used if (i) there are significant trade-offs between fairness and accuracy, and (ii) fairness is deemed significantly more important than accuracy (as portrayed in the value of γ).

Disqualification vs multi-objective optimization (MOO). In spirit, exact disqualification resembles the well studied notion of Pareto efficiency (a classifier should be disqualified if it is Pareto dominated by another classifier in \mathcal{H}); similarly, γ -disqualification resembles further restricting the set of Pareto efficient solutions via an a-priori preference method in which additional information is elicited from a decision-maker (see Chapter 3 in [Hwang and Masud, 2012] for a detailed overview). This is similar in spirit to the value of γ in our work. However, while similar in motivation, on a technical level the concepts of disqualification and Pareto efficiency are distinct. To see this, recall that in our framework the groups S and T are considered symmetric, in the sense that unjustified preferential treatment to either group is considered unfair. This means that "unfairness" cannot be measured in absolute value: for example, two equally accurate classifiers h_S and h_T (where h_g is unfair towards group $g \in \{S, T\}$) may both be on the Pareto frontier. Specifically, in our work the direction of the fairness violation that is pertinent depends on the classifier in question: This breaks the analogy to Pareto efficiency, that like other concepts in MOO, is defined and evaluated w.r.t a fixed set of objective functions.⁵ As a special case, this also explains how our framework is different from the widely used practice of fairness-regularized risk minimization [Kamishima et al., 2012, Zafar et al., 2017, Donini et al., 2018]. Optimizing a regularized objective (a standard loss term combined with a fairness loss term, scaled by some parameter λ) can be viewed as scalarizing a MOO problem, so the question of what exactly is the fairness loss remains. Thus, for similar arguments to those made above, it will not be the case that minimizing a fairness regularized objective will yield a valid solution under our framework.

A second important distinction is that in our work we aim to formalize a framework where a single value γ can capture the trade-offs deemed acceptable between fairness and accuracy; specifically, this value should not depend on the metrics used to quantify what fairness and accuracy mean for a particular task. This is important as indeed fairness and accuracy can mean different things in different contexts, and requires setting up principled translations between units of fairness and units of accuracy, which serves as a fundamental part (and contribution) of our work.

1.2 Our contributions

Our primary contributions are:

1. **Formalizing** γ **-disqualification**, a flexible framework for navigating the space of fairness-accuracy tradeoffs with respect to a benchmark class \mathcal{H} . The framework can be instantiated with (potentially different) loss functions for measuring accuracy and for measuring dispar-

⁴To see this, consider an extreme case in which the data is highly biased, such that all high-accuracy models are in reality extremely unfair. In this situation, it's not necessarily true that we want to insist on returning an accurate model (in fact, if we sufficiently value fairness, we might insist on returning a model that is fair and hence with very low accuracy).

⁵However, our treatment of fair risk minimization in Section 4 reveals that while the two solution concepts are fundamentally distinct, we can use the established work on Pareto efficiency to derive optimal γ -fair classifiers.

ity in predictions, and is parametrized by a "scaling function" that compares differences in accuracy to differences in disparities.

- 2. **Properly scaling fairness and accuracy.** The scaling function is an important ingredient in our framework, and we develop a methodology for choosing it. One aspect we highlight is the minimal level γ for which the Bayes optimal predictor is not γ -disqualified by any other classifier. When possible, we aim to select the scaling function in a way that "anchors" this value at $\gamma=1$. We instantiate this approach for two commonly used accuracy measures. First, for the squared loss, we put forward a natural scaling function, prove that it satisfies the above requirement, and also show that it has several desirable properties in the regime $\gamma<1$. Second, we consider the 0/1 loss and show that it behaves quite differently: there is no "reasonable" scaling function that guarantees the above requirement. These "case studies" are valuable for enabling deployment of our framework, and they also illuminate how different choices of loss functions can quantitatively affect the tension between fairness and accuracy.
- 3. **Fair risk minimization.** We present an algorithm that, given a dataset of labeled examples and the parameter γ , finds an approximately optimal (most accurate) predictor that is not γ —disqualified by another classifier in a given hypothesis class \mathcal{H} . The algorithm is stated as a reduction to the task of approximating the Pareto frontier of \mathcal{H} , where the latter task is well studied, and can be accomplished efficiently for simple classes such as linear regression.

Organization. The rest of this manuscript is organized as follows. In Section 2 we formalize the γ -disqualification framework. Section 3 discusses the question of determining the appropriate "translation" between accuracy and fairness. In Section 4 we study fair risk minimization. Finally, Section 5 discusses example applications, an extension to a "multi-group" setup and further related work. Full proofs are deferred to the appendix.

2 Disqualification

Setup. We use \mathcal{X} to denote the feature space and $A \in \{0,1\}$ group membership, and consider randomized classifiers $h: \mathcal{X} \times A \to \hat{\mathcal{Y}} = [0,1]$, where h(x,g) is the probability that h will predict a label 1 for an individual with features x from group $g \in \{S,T\}$. We use $\ell_A: Y \times \hat{Y} \to [0,1]$ to denote the loss we use to measure global accuracy and $\ell_B: Y \times \hat{Y} \to [0,1]$ to denote the loss we use to measure disparities between groups.

Quantifying unfairness. We say that a classifier is *loss balanced* (for the positive class) w.r.t S and T if the average loss ℓ_B of individuals with y=1 is the same across S and T. We quantify how far a given classifier is from satisfying this requirement, in a specific direction (e.g. how worse-off are the members of S are, relative to the members of S) via the notion of *loss imbalance*:

Definition 2.1 (Loss imbalance). *Fix two groups S*, T *and a loss* ℓ_B . *The* loss imbalance (for the positive class) of a classifier $h: \mathcal{X} \to [0,1]$ in the direction $T \to S$ is

$$dLossImb(h; T \rightarrow S, \ell_B) = \psi(h, S) - \psi(h, T)$$

where
$$\psi(h, C) = \mathbf{E}_{x,g,y \sim \mathbb{P}}[\ell(h(x), y) | y = 1, g = C].$$

We note that our notion of loss imbalance (Definition 2.1) can be viewed as quantifying unfairness, where the notion of fairness generalizes many common existing group fairness definitions. In particular, when ℓ_B is the (expected) 0/1 loss, it recovers both *Balance for the positive class* [Kleinberg et al., 2016] and *Equal Opportunity* [Hardt et al., 2016], depending on whether the classifiers in questions are binary or randomized. See Appendix B for details. For the rest of this paper we consider randomized classifiers and think of ℓ_B as the expected 0/1 loss.

Formalizing γ -disqualification. We want to say that a classifier h' disqualifies a second classifier h if the former improves the loss imbalance (as measured w.r.t ℓ_B) over the latter, without hurting global accuracy (as measured w.r.t ℓ_A) too much – with the exact tradeoff specified by the parameter γ . However, as we discuss in Section 3, naively comparing the difference in the two quantities is an "apples to oranges" comparison. We address this by applying a scaling function to the difference in accuracy before comparing it to the difference in imbalance. In principle, the scaling is allowed to depend on γ , so the scaling function is a mapping f from $\mathbb{R} \times (0, \infty)$ to \mathbb{R} . For a fixed value $\gamma \in (0, \infty)$, we use f_{γ} to denote the resulting function $f_{\gamma} : \mathbb{R} \to \mathbb{R}$.

Definition (γ -Disqualification). A classifier h' γ -disqualifies a classifier h w.r.t losses ℓ_A and ℓ_B and a scaling function f if

$$dLossImb(h; \ell_B) - dLossImb(h'; \ell_B) > f_{\gamma}(\max\{\ell_A(h') - \ell_A(h), 0\})$$
(1)

where $dLossImb(\cdot)$ is computed in the direction for which $dLossImb(h; \ell_B) \geq 0$.

Note that to assess whether h' disqualifies h, we first take into account the direction of the fairness violation of the original classifier, h. This is designed to ensure that if h is unfair e.g. because it favours the members of T over S (i.e. the imbalance of h is positive in the direction $T \to S$; see Definition 2.1), only a classifier that improves the imbalance *in this direction* can disqualify h.

"Consider the alternatives": γ -Fairness. Finally, we refer to the requirement that a classifier not be disqualified by any alternative in a class \mathcal{H} of alternative classifiers as γ -Fairness:

Definition (γ -Fairness). Fixing loss functions ℓ_A and ℓ_B and a scaling function f_{γ} , we say that a classifier h satisfies γ -fairness w.r.t $\mathcal{H} \subseteq \hat{Y}^{X \times A}$ if no classifier $h' \in \mathcal{H}$ γ -disqualifies h w.r.t ℓ_A , ℓ_B and f_{γ} . When \mathcal{H} is unconstrained, $\mathcal{H} = \hat{Y}^{X \times A}$, we simply say that h is γ -fair.

3 Specifying the scaling function

Given losses ℓ_A and ℓ_B for measuring accuracy and fairness (respectively), how should we choose the scaling function that "translates" units of accuracy into units of fairness? As a warm-up, we revisit the simple examples of Figure 1. Recall that there we were concerned with binary classifiers, and imbalance was measured as the disparity in TPR across S and T (so ℓ_A and ℓ_B are both the 0/1 loss). Relative to h_w , the classifier h_{w_a} improves fairness by 0.5 and hurts accuracy by 0.25 – but how should we think about comparing between these numbers? One natural perspective is to consider the (worst-case) impact a single individual can have on both accuracy and fairness. In

this case, even though $\ell_A = \ell_B$, the (potential) contribution of a single individual to the change in unfairness is much greater than it is to the change in accuracy loss, since the latter is measured globally over the entire population whereas fairness conditions on both group membership and label. This means that at the very minimum, the scaling function we choose must take this into account by appropriately "up-weighting" accuracy before comparing it to the fairness.

To make this intuition more precise (and to generalize to arbitrary choices of ℓ_A and ℓ_B), consider the following hypothetical question: How much can a certain improvement in accuracy – resulting from changing the prediction of a single individual – impact unfairness, in the worst case? To make this concrete, consider some classifier h and individual x, and define a new classifier h' as follows:

$$h'(\tilde{x}) = \begin{cases} h^{\star}(\tilde{x}) & \tilde{x} = x \\ h(\tilde{x}) & \tilde{x} \neq x \end{cases}$$

That is, h' is identical to h, except we switched the prediction of x from h(x) to $h^*(\tilde{x})$, which we use to denote the Bayes optimal prediction for \tilde{x} according to ℓ_A . Recall that in general, the Bayes optimal predictor h^* is maximally accurate (w.r.t ℓ_A), but potentially unfair (w.r.t ℓ_B). Thus, in comparing h' to h, accuracy has improved (say by $\tau_A(x)$) while fairness has potentially deteriorated (say by $\tau_B(x)$). Does h disqualify h'? According to Equation (1), the original classifier h γ -disqualifies h' if the scaling function f_{γ} is such that

$$\tau_B(x) > f_{\gamma}(\tau_A(x)).$$

Generalizing this argument to arbitrary choices of the initial classifier h and individual x, we conclude that w.r.t some fixed scaling function f_{γ} , the Bayes optimal predictor h^* will be γ^* -fair, where $\gamma^* \leq \infty$ is the smallest possible value that guarantees that for any initial classifier h, the maximal deterioration in fairness $\tau_B(x)$ possible from some improvement in accuracy $\tau_A(x)$ does not exceed $f_{\gamma^*}(\tau_A(x))$. Given a scaling function f_{γ} , the value of γ^* provides a threshold, where improving accuracy (moving individuals to the Bayes optimal prediction) will not lead to disqualification (even though it might degrade fairness).

The above discussion leads to an important principle in choosing the translation function: whenever possible, we want the Bayes optimal predictor to be fair with $\gamma^*=1$. By "anchoring" the value of γ^* to a fixed value, such as 1, we guarantee that the extent to which h^* is fair is invariant under scalar multiplication of the loss function ℓ_A . We view this as a desirable property: the accuracy and fairness losses are (initially) incomparable, switching from measuring accuracy using ℓ_A to measuring accuracy using $100 \cdot \ell_A$ does not meaningfully change the fairness-accuracy trade-offs in question – which is precisely what our notion of fairness is aimed at capturing.

Thus, given a pair of losses (ℓ_A, ℓ_B) , we aim to choose the scaling function in a way that guarantees that the Bayes optimal classifier according to ℓ_A will be γ -fair with $\gamma=1$. In this section, we fix ℓ_B as the 0/1 loss and instantiate this approach w.r.t two common loss functions for measuring accuracy: the squared loss and the 0/1 loss. For the squared loss we put forward a natural scaling function, prove that it satisfies the above requirement, and also show that it has several desirable properties in the regime $\gamma<1$. On the other hand, for the 0/1 loss, we show that there is no "reasonable" scaling function that guarantees the Bayes optimal classifier is 1—fair. This is because

the loss is (in a certain sense) highly "non-Lipschitz": achieving tiny accuracy improvements can require an unbounded degradation in fairness. This negative result highlights that, for these loss functions, fairness and accuracy might be wildly conflicting.

3.1 Scaling w.r.t the squared loss

For the squared loss $\ell_A(y,\hat{y}) = (y - \hat{y})^2$ we consider the following scaling function:

$$f_{\gamma}(a) = \sqrt{\gamma \cdot \frac{2a}{\eta}}, \quad \text{where } \eta = \min_{g \in \{S,T\}} \Pr_{x,y}[x \in g \land y = 1]$$
 (2)

The scaling function first applies a linear scaling to account for the fact that fairness and accuracy are computed by aggregating over different-sized sets (akin to the "warm-up" discussion above), and then applies a square root to account for the discrepancy between ℓ_A and ℓ_B . We prove that it indeed satisfies the principle described above:

Theorem 1. When ℓ_B is the expected 0-1 loss and ℓ_A is the squared loss, the Bayes optimal predictor w.r.t ℓ_A , $h^*(x,g) = \mathbf{E}_{\mathbb{P}}[y|x,g]$ is γ -fair for $\gamma = 1$ w.r.t the scaling function defined in Equation (2).

See Appendix C for the proof. For intuition, we note that one way in which we can always transform h^* into a perfectly balanced predictor is to increase the predictions for everyone in S by $\varepsilon \triangleq Imb(h^*)$. This transformation increases the squared loss by exactly ε^2 (scaled to the relative mass of S). Thus, for this new classifier to not disqualify h^* when $\gamma=1$, the scaling function should be chosen such that $\varepsilon \leq f_1(\varepsilon^2)$.

We note that the principle described above does not fully constrain the form of the scaling function, since it only specifies a constraint for $\gamma=1$. We additionally prove that our choice of the scaling function given in Equation (2) (and specifically, the incorporation of γ under the square root) guarantees another desirable property: that convex combinations of the Bayes optimal predictor (which is fair for $\gamma=1$) and the optimal (most accurate) perfectly balanced predictor (which is fair for $\gamma=0$) are also fair (for the class of convex combinations of these two classifiers, and for γ specified by the convex combination):

Theorem 2. Let $h_1 = \mathbf{E}_{\mathbb{P}}[y|x,g]$ and h_0 denote the optimal γ -fair classifiers for $\gamma = 1$ and $\gamma = 0$, respectively. Furthermore, define $\tilde{\mathcal{H}}$ as the collection of all convex combinations of some 1-fair and some 0-fair classifiers. Then,

- (i) The classifier $h_{\gamma}=\gamma\cdot h_1+(1-\gamma)\cdot h_0\in \tilde{\mathcal{H}}$ is $(\gamma,\tilde{\mathcal{H}})$ -fair, but
- (ii) There is a 0-fair h_0' such that the classifier $h_\gamma' = \gamma \cdot h_1 + (1-\gamma) \cdot h_0' \in \tilde{\mathcal{H}}$ is not $(\gamma, \tilde{\mathcal{H}})$ -fair.

See Appendix D for the proof. Theorem 2 gives some intuition for fairness in the regime $0 < \gamma < 1$, and crucially relies on our choice of scaling function (and in particular on the way that γ affects the scaling). The first part of Theorem 2 highlights that in this case, $\gamma \in (0,1)$ serves as a "knob" that, as it shrinks, brings down the imbalance level – from $dImb(h_1)$ (which is too high when $\gamma < 1$), to a sufficiently low level (here, $\gamma \cdot dImb(h_1)$). However, the second part of the theorem highlights that there is more to our fairness requirement than simply reaching a sufficiently low level of imbalance: Indeed, h_{γ} and h'_{γ} are both *equally imbalanced*, yet only the first is fair. This is in line with our initial motivation, in which we called for a more relative (context-dependent) perspective, that considers the alternatives and not only the objective level of balance (or imbalance).

3.2 Scaling w.r.t the 0/1 loss

Interestingly, not every combination of losses ℓ_A and ℓ_B has a scaling function that can guarantee the Bayes optimal classifier is always fair for $\gamma=1$. We prove this is the case when we switch ℓ_A from the squared loss to the expected 0/1 loss.

Theorem 3. Fix ℓ_B and ℓ_A to be the expected 0/1 loss. Then there is no "reasonable" scaling function f_{γ} that guarantees that the Bayes optimal classifier for the 0/1 loss is γ -fair for $\gamma = 1$.

Here, the requirement on the scaling function f_{γ} is minimal: we only require that f_{γ} does not "blow up" when $\gamma < \infty$. See Appendix E for the proof. The intuition for Theorem 3 is that the binary nature of the Bayes optimal classifier for the expected 0-1 loss can work to amplify even very small differences between groups, in a way that our disqualification framework considers unfair (unless $\gamma = \infty$). To see this, consider the simple case in which there is no information except group membership. In this case, the Bayes optimal classifier predicts 0.0 for an individual from a group whose base rate is below 0.5, and 1.0 for an individual from a group whose base rate is above 0.5. This has the effect of *amplifying* differences between the two groups: even a very small difference in base rates (e.g., 0.49 for S and 0.51 for T) translates to a maximal imbalance of 1.0. In this case, our framework views such a classifier as unfair, because e.g. a classifier that predicts 0.5 for everyone will disqualify the Bayes optimal classifier: it maximally improves the imbalance (from 1.0 to 0.0) at a minimal cost to accuracy.

4 Fair risk minimization

While \mathcal{H} itself doesn't necessarily contain a classifier that is γ -fair w.r.t \mathcal{H} , the class of convex combinations of classifiers from \mathcal{H} , denoted $\Delta(\mathcal{H})$, provably does (see Appendix F for the proof). Thus, thinking of γ -fairness as a hard constraint, a natural objective is to output the most accurate classifier in $\Delta(\mathcal{H})$ that is γ -fair w.r.t \mathcal{H} . We refer to this as the *fair risk minimization* problem. In practice, however, we will typically be interested in the empirical counterpart of this problem (or the *fair ERM*), where all the quantities are computed w.r.t a finite sample. To accommodate the transition from working with the underlying distribution \mathbb{P} to working with an i.i.d. sample $D \sim \mathbb{P}^m$, we define a natural notion of *approximate* disqualification, and prove that the resulting approximate fairness definition implied by it generalizes from a sample to the underlying distribution. Intuitively, a classifier $h'(\alpha, \gamma)$ -disqualifies h if the requirement in Equation (1) holds, even when we add an additive slack term α to both the difference in imbalance and the difference in loss.

In this section we prove that while γ -fairness and Pareto efficiency are distinct (as solution concepts), the latter is sufficiently informative for the purposes of fair risk minimization. Specifically, we show that given oracle access to an approximation of the Pareto frontier of \mathcal{H} w.r.t accuracy and imbalance (in a *fixed* direction, say $T \to S$) we can solve the fair ERM problem.

Theorem 4. Fix a class \mathcal{H} , tradeoff parameter γ , approximation parameter α and a dataset D. There exists an algorithm that, for every $\varepsilon \leq \alpha$, produces a classifier h that is (i) (α, γ) -fair w.r.t \mathcal{H} on D, (ii) at least as accurate (on D) as the optimal $(\alpha - \varepsilon, \gamma)$ -fair classifier in \mathcal{H} . Furthermore, the algorithm runs in

time $poly(1/\epsilon)$ and makes $poly(1/\epsilon)$ queries to $PF^{\Delta(\mathcal{H})}$, where $PF^{\Delta(\mathcal{H})}(\tau)$ returns a classifier from $\Delta(\mathcal{H})$ whose accuracy is optimal given that its imbalance level doesn't exceed $\tau \in [-1,1]$.

The algorithm itself is simple: it iterates through the classifiers in the Pareto frontier of \mathcal{H} to find the most accurate one that is not empirically (approximately) disqualified by another classifier on the Pareto frontier. Using the property of Pareto efficiency, this suffices for the desired guarantee; See Appendix F for the full proof.

Theorem 4 demonstrates that the fair ERM problem is no harder than the problem of approximately computing the Pareto frontier of \mathcal{H} . In some simple cases, the later can be solved efficiently. For example, when \mathcal{H} is the class of linear classifiers with bounded norm over \mathbb{R}^d , the overall running time of A will be $\operatorname{poly}(1/\varepsilon,d)$ (since every query to $\operatorname{PF}(\tau;\mathcal{H})$ can be obtained as the solution to a convex program with d variables). While this isn't true for general classes, estimating the entire Pareto frontier is a well studied task with a variety of existing algorithms and heuristics [Fliege and Svaiter, 2000].

Remark. We note that the algorithm in Theorem 4 only approximately solves the fair ERM problem, in the sense that it produces a (α, γ) -fair classifier whose sample-accuracy is only competitive with the sample-accuracy of the best $(\alpha - \varepsilon, \gamma)$ -fair classifier. Without assuming anything about \mathcal{H} , this gap could be substantial. This highlights that the approximation guarantee is closely related to the Lipschitzness of the Pareto frontier of \mathcal{H} ; see Appendix F for an additional discussion.

5 Discussion

Applications. Our framework can be used to select from (and compare between) different learning strategies, with and without an external quantification of the appropriate trade-off parameter γ . For a concrete example, suppose a data-scientist fits a standard learning algorithm to their data, say the Adult Income dataset, yielding a predictor h_{ERM} whose squared error on the test set is, say, 0.149. Mindful of fairness, they also evaluate the difference in the true positive rate between men and women, finding it to be significant: under h_{ERM} , men with a positive label receive scores that are on average 0.11 higher than women with positive labels. In response, they consider a different strategy, such as a adding an explicit fairness regularization penalty to the ERM objective. This yields a second classifier $h_{fairERM}$, whose squared error on the test set is now 0.15 but whose imbalance has dropped to 0.051. What should guide the comparison (and eventual choice) between h_{ERM} and $h_{fairERM}$? Our framework provides a simple and clear criteria: we should pick the classifier that improves fairness only if for us, fairness is $20 \times$ as important as accuracy is.⁶ An important aspect of our framework is that the value of γ can be elicited once from an external party such as a regulator, and this can be applied by data scientists to select classifiers for a variety of different tasks (with the appropriate scaling functions). We also note that the same principles can be used even in the absence of an external quantification of γ . Fixing a benchmark class \mathcal{H} and a learning strategy that results in a classifier h, we can solve for the minimal value $\hat{\gamma}$ for which there exists $h' \in \mathcal{H}$ that $\hat{\gamma}$ -disqualifies h. The resulting value $\hat{\gamma}$ can be thought of as the "effective unfairness"

⁶By definition, $h_{fairERM}$ γ -disqualifies h_{ERM} if the normalized improvement in fairness exceeds the loss to accuracy: $0.11-0.051>f_{\gamma}(0.15-0.149)$. The Adult Income dataset is imbalanced, and so in this case $\eta=\Pr[\text{is}-\text{woman} \land y=1]\approx 0.03$. The scaling function f from Theorem 1 therefore "re-scales" the loss difference as $f_{\gamma}(0.001)=\sqrt{2\gamma\cdot0.001/0.03}$. Solving for γ , we obtain that the $h_{fairERM}$ γ -disqualifies h_{ERM} only when $\gamma<0.05$.

of the algorithm w.r.t the benchmark, and can serve as additional metric, that is both simple and interpretable, in comparing different learning algorithms.

Disqualification in the "multi-group" setup. While violations of "group fairness" notions (such as an imbalance between S and T) can serve as meaningful red flags, focusing on aggregate behavior over the entire group can open the door to abuses targeting sub-populations. Individual fairness, pioneered by [Dwork et al., 2012], provides strong fairness protections but requires a task-specific similarity metric, which can be challenging to obtain. A more recent line of work, starting with [Hébert-Johnson et al., 2017, Kearns et al., 2018], strengthens group fairness by requiring its guarantee to hold for every group in a large collections \mathcal{C} of overlapping sub-populations. Such "multi-group" fairness notions can be instantiated with different fairness measures: Multicalibration [Hébert-Johnson et al., 2017] requires that the risk scores be calibrated on every group in the collection whereas Kearns et al. [Kearns et al., 2018] suggest imposing demographic parity between groups in the collection. We can thus similarly extend our framework to the multi-group setup by requiring that for for every pair of groups $S,T\in\mathcal{C}$ (where \mathcal{C} is a large collection of overlapping sets), no classifier γ -disqualifies the given classifier where the imbalance measures the predictions of S vs T. The tension between fairness and accuracy is an important issue in the study of multi-group fairness, as it directly impacts our ability to strengthen the fairness guarantee via enriching the collection C. For fairness notions that are not in tension with accuracy, such as calibration, there is no inherent accuracy downside to enriching \mathcal{C} (note, however, that guaranteeing fairness for richer collections might require higher running time or sample complexity). In particular, the Bayes-optimal predictor satisfies multi-calibration for any collection of sets [Hébert-Johnson et al., 2017]. On the other hand, for notions like multi-demographic parity, enriching the collection of sets might make informative predictions impossible: e.g. if the collection includes many sets with different base rates. Thus, whereas it would be quite desirable to obtain multicalibration with respect to entire computation classes (e.g. all the sufficiently-large sets that can be efficiently identified from the data), for multi-demographic parity or multi-balance [Kearns et al., 2018] this might make very little sense: the collection of sets should be carefully tailored to the problem at hand, and adding additional sets might "over-constrain" the problem, forcing the predictions to be uninformative. In this respect, multi-fairness with $\gamma = 1$ can be considered as an alternative to the multi-balance notion from [Kearns et al., 2018] (which is equivalent to multi-fairness with $\gamma = 0$): the fairness guarantee is more relaxed, but the collection of sets can be enriched almost arbitrarily without making the definition overly restrictive (a corollary of Theorem 1).

Additional related work. The notion of Equal Opportunity was proposed in the seminal work of [Hardt et al., 2016], and the related notion of balance, which we focus on here, was proposed in [Kleinberg et al., 2016]. The impossibilities results for simultaneously obtaining calibration (which can be viewed as a minimal accuracy requirement) and error rate parity were established in [Kleinberg et al., 2016, Chouldechova, 2017]. They were strengthened by [Pleiss et al., 2017] who show that even a single parity constraint (e.g. equalizing false positive rates) can't be meaningfully obtained together with calibration. The fairness community has taken several approaches in light of these basic trade-offs. One approach seeks to understand the conditions under which fairness is desirable despite (or in spite) of its cost to accuracy, such as when the data itself is discriminatory [Blum and Stangl, 2019] or when the fairness dynamics require it [Jung et al., 2020]. Different

works attempt to quantify fairness in a way that stands in less stark opposition to accuracy to begin with [Hébert-Johnson et al., 2017, Kim et al., 2019, Blum and Lykouris, 2019, Kallus and Zhou, 2019, Rothblum and Yona, 2021], or to alleviate the trade-offs by gathering more informative data [Chen et al., 2018, Garg et al., 2019].

References

- [Barocas and Selbst, 2016] Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671.
- [Blum and Lykouris, 2019] Blum, A. and Lykouris, T. (2019). Advancing subgroup fairness via sleeping experts. *arXiv preprint arXiv:1909.08375*.
- [Blum and Stangl, 2019] Blum, A. and Stangl, K. (2019). Recovering from biased data: Can fairness constraints improve accuracy? *arXiv* preprint arXiv:1912.01094.
- [Chen et al., 2018] Chen, I., Johansson, F. D., and Sontag, D. (2018). Why is my classifier discriminatory? In *Advances in Neural Information Processing Systems*, pages 3539–3550.
- [Chouldechova, 2017] Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163.
- [Donini et al., 2018] Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J., and Pontil, M. (2018). Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*.
- [Dwork et al., 2012] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.
- [Fliege and Svaiter, 2000] Fliege, J. and Svaiter, B. F. (2000). Steepest descent methods for multi-criteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494.
- [Garg et al., 2019] Garg, S., Kim, M. P., and Reingold, O. (2019). Tracking and improving information in the service of fairness. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 809–824.
- [Hardt et al., 2016] Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.
- [Hébert-Johnson et al., 2017] Hébert-Johnson, U., Kim, M. P., Reingold, O., and Rothblum, G. N. (2017). Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*.
- [Hwang and Masud, 2012] Hwang, C.-L. and Masud, A. S. M. (2012). *Multiple objective decision making—methods and applications: a state-of-the-art survey*, volume 164. Springer Science & Business Media.
- [Jung et al., 2020] Jung, C., Kannan, S., Lee, C., Pai, M., Roth, A., and Vohra, R. (2020). Fair prediction with endogenous behavior. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 677–678.

- [Kallus and Zhou, 2019] Kallus, N. and Zhou, A. (2019). The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In *Advances in Neural Information Processing Systems*, pages 3433–3443.
- [Kamishima et al., 2012] Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 35–50. Springer.
- [Kearns et al., 2018] Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerry-mandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2564–2572.
- [Kim et al., 2019] Kim, M. P., Ghorbani, A., and Zou, J. (2019). Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI*, *Ethics, and Society*, pages 247–254.
- [Kleinberg et al., 2016] Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
- [Kusner et al., 2017] Kusner, M. J., Loftus, J., Russell, C., and Silva, R. (2017). Counterfactual fairness. In *Advances in neural information processing systems*, pages 4066–4076.
- [Pleiss et al., 2017] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689.
- [Rothblum and Yona, 2021] Rothblum, G. N. and Yona, G. (2021). Multi-group agnostic pac learnability. *arXiv preprint arXiv:2105.09989*.
- [Zafar et al., 2017] Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.

A Additional notation

We use \mathcal{X} to denote the space of covariates, A a binary group membership indicator (defining two groups, which we denote by T and S), and $\mathcal{Y} = \{0,1\}$. We assume an underlying (but unknown) distribution \mathbb{P} on $\mathcal{X} \times A \times \mathcal{Y}$. W will sometimes use D_g to denote \mathbb{P} restricted to samples from group $g \in \{S, T\}$. Additionally, μ_g denotes the fraction of individuals that are in g (with $\mu_S = 1 - \mu_T$) and β_g denotes the base rate for g: $\beta_g \equiv \mathbf{Pr}_{x,y \sim D_g}[y = 1]$.

A binary classifier is a mapping $h: \mathcal{X} \times A \to \mathcal{Y}$. A hypothesis class \mathcal{H} is a collection of such binary classifiers. Since we will generally work with convex combinations of binary classifiers, we will consider randomized classifiers $h: \mathcal{X} \times A \to \hat{\mathcal{Y}}$, where $\hat{\mathcal{Y}} = [0,1]$ and h(x,g) is the probability that h labels an individual from g with covariates x as positive.

A loss function ℓ is mapping from $Y \times \hat{Y} \to \mathbb{R}$. Slightly abusing notation, we use $\ell_{\mathbb{P}}(h) = \mathbb{E}_{x,g,y\sim\mathbb{P}}[\ell(h(x,g),y)]$ to denote the loss of a randomized classifier h w.r.t the distribution \mathbb{P} , and we will drop the subscript \mathbb{P} when it is clear from context. Given a loss ℓ , we denote by h_{ℓ}^{\star} the Bayes optimal classifier w.r.t ℓ : $h_{\ell}^{\star} \in \arg\min_{h:\mathcal{X} \times A \to \hat{\mathcal{Y}}} \ell(h)$.

In this work we will focus on two popular choices: the expected zero-one loss (the probability that *h* assigns the correct binary label) and the squared loss:

$$\ell^{0-1}(\hat{y}, y) = \hat{y} \cdot (1 - y) + (1 - \hat{y}) \cdot y$$
$$\ell^{2}(\hat{y}, y) = (\hat{y} - y)^{2}$$

The Bayes optimal classifiers for these loss functions are:

$$h_{\ell^{2}}^{\star}(x,g) = \underset{x,g,y \sim \mathbb{P}}{\mathbf{E}} \left[y | x, g \right],$$

$$h_{\ell^{0-1}}^{\star}(x,g) = \mathbf{1} \left[\underset{x,g,y \sim \mathbb{P}}{\mathbf{E}} \left[y | x, g \right] \ge \frac{1}{2} \right] = \mathbf{1} \left[h_{\ell^{2}}^{\star}(x,g) \ge \frac{1}{2} \right]$$

B Loss imbalance generalizes existing notions

Our starting point is a definition of *group fairness* that seeks to equalize some quantity $\psi(h, C)$ across groups S and T. Assuming higher values of $\psi(h, C)$ are better, we can then define the (directed) degree of unfairness as

$$dUnfairness(h; \psi, C \to C') = \psi(h, C) - \psi(h, C')$$
(3)

Such that dUnfairness(h; ψ , $T \to S$) measures the extent to which T receives favourable outcomes (in the sense implied by ψ). For example,

- $\psi^{DP}(h,C) = \mathbf{E}_{x,g,y\sim\mathbb{P}}[h(x) \mid g=C]$ recovers the definition of Demographic Parity,
- $\psi^{PosBalance}(h,g) = \mathbf{E}_{x,g,y\sim\mathbb{P}}[h(x) | y = 1, g = C]$ recovers the definition of balance for the positive class [Kleinberg et al., 2016], and, when h is binary, also the definition of Equal Opportunity [Hardt et al., 2016]

Remark. While this is not the focus of our work, we can generalize this formulation to situations in which by fairness we mean equalizing multiple quantities. For example, in Equal Odds [Hardt et al., 2016], both $\psi_1(h,g) = \text{FPR}(h,g)$ and $\psi_2(h,g) = \text{TPR}(h,g)$ should be equalized across groups; in this case, the directed unfairness can be defined more generally as

$$dUnfairness(h; \psi, C \to C') = \max \{ \psi_1(h, C') - \psi_1(h, C), \ \psi_2(h, C) - \psi_2(h, C') \}$$
(4)

B.1 Imbalance and loss imbalance

In this work, our focus will be on taking fairness to mean balance for the positive class. We refer to the directed unfairness dUnfairness $(h; \psi^{PosBalance}, T \to S)$ more simply as the directed imbalance, denoted $dImb(h; T \to S)$. In fact, we will be working with a slightly more general notion of imbalance, which we refer to as loss imbalance. Instead of directly comparing the expected scores of the positive members of T with expected scores of the positives members of S, loss imbalance compares the difference in their expected losses.

Definition B.1 (Loss Imbalance). *Given a loss* $\ell : \hat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$, the loss imbalance of h in the direction $T \to S$ is defined as

$$dLossImb(h; T \to S, \ell) = \psi^{PosLossBalance}(h, S) - \psi^{PosLossBalance}(h, T)$$
 (5)

where
$$\psi^{PosLossBalance}(h,C) = \mathbf{E}_{x,g,y\sim\mathbb{P}}[\ell(h(x),y) \mid y=1,g=C].$$

Note the change in order – subtracting T from S, as opposed to S from T in Equation (3). This is because for loss ℓ , lower is better.

Loss imbalance is a generalization of the notion of imbalance. The next lemma shows that imbalance is simply the loss imbalance as measured w.r.t the expected zero one loss.

Lemma B.2. For any
$$h: \mathcal{X} \to \hat{\mathcal{Y}}$$
, $dLossImb(h; \ell^{0-1}) = dImb(h)$.

Proof. Recall that the 0-1 loss is defined as $\ell^{0-1}(h(x),y) = h(x) \cdot (1-y) + (1-h(x)) \cdot y$, so $\ell^{0-1}(h(x),1) = 1-h(x)$. We therefore have

$$\begin{split} LossImb(h;\ell^{0-1}) &= \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[\ell(h(x),y) \mid y=1,g=S] - \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[\ell(h(x),y) \mid y=1,g=T] \\ &= \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[\ell(h(x),1) \mid y=1,g=S] - \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[\ell(h(x),1) \mid y=1,g=T] \\ &= \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[1-h(x) \mid y=1,g=S] - \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[1-h(x) \mid y=1,g=T] \\ &= \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[h(x) \mid y=1,g=T] - \mathop{\mathbf{E}}_{x,g,y\sim\mathbb{P}}[h(x) \mid y=1,g=S] \\ &= Imb(h) \end{split}$$

C Proof of Theorem 1

C.1 Warmup: A simple case

For simplicity, let ℓ denote the squared loss. We want to show a choice of $f = f_1$ for which no classifier can disqualify $h_{\ell^2}^{\star}$:

$$\forall h': \mathcal{X} \to [0,1]: \quad Imb(h_{\ell^2}^*) - Imb(h') \le f(\min\{0, \ell(h') - \ell(h_{\ell^2}^*)\})$$

Since $h_{\ell^2}^{\star}$ is optimal, this is the same as

$$\forall h': \mathcal{X} \rightarrow [0,1]: \quad Imb(h_{\ell^2}^{\star}) - Imb(h') \leq f(\ell(h') - \ell(h_{\ell^2}^{\star}))$$

We will show this can be done in the simple case in which $\mathcal{X} = \emptyset$, and $h_{\ell^2}^*$ predicts β_T for T and β_S for S, and has an imbalance of $\beta_T - \beta_S$. There are two ways for a classifier to improve the imbalance: increase the prediction for S or decrease the prediction for T. Consider the first case, so h' is identical to $h_{\ell^2}^*$, but adds ε to the prediction of S. By definition, the imbalance improves by ε . The increase in loss, on the other-hand, is proportional to ε^2 :

$$\ell(h') - \ell(h_{\ell^2}^{\star}) = d(h', h_{\ell^2}^{\star}) - d(h_{\ell^2}^{\star}, h_{\ell^2}^{\star})$$

$$= d(h', h_{\ell^2}^{\star})$$

$$= u_S \cdot \varepsilon^2$$

Where $d(p, p') = \mathbf{E}_x[(p(x) - p'(x))^2]$. This uses the fact that for every classifier h, the squared loss is related to the Euclidean distance from $h_{\ell^2}^{\star}$, as follows: $\ell(h) = d(h, h_{\ell^2}^{\star}) + t(h_{\ell^2}^{\star})$; Indeed:

$$\ell(h) = \underset{x,y}{\mathbf{E}}[(y - h(x))^{2}]$$

$$= \underset{x}{\mathbf{E}}[h_{\ell^{2}}^{\star}(x) \cdot (1 - h(x))^{2} + (1 - h_{\ell^{2}}^{\star}(x)) \cdot h(x)^{2}]$$

$$= \underset{x}{\mathbf{E}}[h_{\ell^{2}}^{\star}(x) - 2h_{\ell^{2}}^{\star}(x)h(x) + h(x)^{2}]$$

$$= \underset{x}{\mathbf{E}}h_{\ell^{2}}^{\star}(x) - h_{\ell^{2}}^{\star}(x)^{2}] + \underset{x}{\mathbf{E}}[(h_{\ell^{2}}^{\star}(x) - h(x))^{2}]$$

$$= d(h, h_{\ell^{2}}^{\star}) + t(h_{\ell^{2}}^{\star})$$

We can now see that h' doesn't 1—disqualify $h_{\ell^2}^*$:

$$f_1(\ell(h') - \ell(h_{\ell^2}^{\star})) = t_1(1) \cdot t_2(\mu_S \cdot \varepsilon^2) = t_2(\mu_S \cdot \varepsilon^2) = \sqrt{\frac{\mu_S \cdot \varepsilon^2}{\mu \cdot \beta}} \ge \varepsilon = Imb(h_{\ell^2}^{\star}) - Imb(h')$$

as required.

C.2 Full proof of Theorem 1

Proof. For simplicity, denote $h^* \triangleq h_{\ell^2}^*$. Assume h^* is not perfectly balanced (otherwise we are done) and w.l.o.g that the imbalance is in favor of T, so when we write Imb we mean the imabalance in the direction $T \to S$.

We want to prove that for the specified scaling function, no other classifier h 1—disqualifies h^* :

$$f_1(\ell(h) - \ell(h^*)) \ge d\operatorname{Imb}(h^*) - d\operatorname{Imb}(h) \tag{6}$$

For simplicity, we will sometimes write $f_1 = f$. Note that we can indeed write $f(\ell(h) - \ell(h^*))$ (as opposed to $f(\max\{0, \ell(h) - \ell(h^*)\})$), as the definition states) because from the optimality of h^* w.r.t $\ell = \ell^2$ we have that $\ell(h) - \ell(h^*) \ge 0$.

Let *h* be any classifier; define the following quantities for $x \in \mathcal{X}, g \in \{S, T\}$:

$$\Delta_{x,g} \equiv h^{\star}(x,g) - h(x,g)$$

$$m_{x,g} \equiv \Pr_{D_g}(X = x)$$

$$m_{x,g}^{y} \equiv \Pr_{D_g}(X = x|Y = y)$$

Note that for a group g and $y \in \{0,1\}$, both define legal probability measures (they are non-negative, and sum to 1: $\sum_{x \in \mathcal{X}} m_{x,g} = \sum_{x \in \mathcal{X}} m_{x,g}^y = 1$). We can now express both the difference in loss and the difference in imbalance between h^* and h in terms of m, m^1 and Δ , as follows:

$$Imb(h^*) - Imb(h) = \sum_{x} m_{x,T}^1 \cdot \Delta(x,T) - \sum_{x} m_{x,S}^1 \cdot \Delta(x,S)$$
 (7)

$$\ell(h) - \ell(h^*) = \mu_T \cdot \sum_{x} m_{x,T} \cdot \Delta(x,T)^2 + \mu_S \cdot \sum_{x} m_{x,S} \cdot \Delta(x,S)^2$$
 (8)

For (7), we first note that for any classifier \tilde{h} ,

$$Imb(\tilde{h}) = \mathbf{E}[\tilde{h}(x,T)|y=1] - \mathbf{E}[\tilde{h}(x,S)|y=1] = \sum_{x} m_{x,T}^1 \cdot h(x,T) - \sum_{x} m_{x,S}^1 \cdot h(x,S)$$

So

$$\begin{split} Imb(h^{\star}) - Imb(h) &= \left(\sum_{x} m_{x,T}^{1} \cdot h_{1}(x,T) - \sum_{x} m_{x,S}^{1} \cdot h_{1}(x,S) \right) - \left(\sum_{x} m_{x,T}^{1} \cdot h(x,T) - \sum_{x} m_{x,S}^{1} \cdot h(x,S) \right) \\ &= \sum_{x} m_{x,T}^{1} \cdot \Delta(x,T) - \sum_{x} m_{x,S}^{1} \cdot \Delta(x,S) \end{split}$$

For (8), note that the loss of a classifier \tilde{h} restricted to a group g can be written as

$$\ell_{g}(\tilde{h}) = \sum_{x} m_{x,g} \cdot \left(h^{\star}(x,g) \cdot (\tilde{h}(x,g) - 1)^{2} + (1 - h^{\star}(x,g) \cdot \tilde{h}(x,g)^{2} \right)$$
$$= \sum_{x} m_{x,g} \cdot \left(h^{\star}(x,g) - 2h^{\star}(x,g)\tilde{h}(x,g) + \tilde{h}(x,g)^{2} \right)$$

By direct calculation, this implies that

$$\ell_g(h) - \ell_g(h^*) = \sum_x m_{x,g} \cdot (h(x,g) - h^*(x,g))^2 = \sum_x m_{x,g} \cdot \Delta(x,g)$$

From which (8) follows, since $\ell(h) - \ell(h^*) = \mu_T \cdot [\ell_T(h) - \ell_T(h^*)] + \mu_S \cdot [\ell_S(h) - \ell_S(h^*)]$. We can now obtain the required:

$$f_1\left[\ell(h) - \ell(h^*)\right] = \sqrt{\frac{2\left[\mu_T \cdot \sum_x m_{x,T} \cdot \Delta(x,T)^2 + \mu_S \cdot \sum_x m_{x,S} \cdot \Delta(x,S)^2\right]}{\mu\beta}}$$
(9)

$$\geq \frac{\sqrt{\mu_T \cdot \sum_x m_{x,T} \cdot \Delta(x,T)^2} + \sqrt{\mu_S \cdot \sum_x m_{x,S} \cdot \Delta(x,S)^2}}{\sqrt{\mu\beta}} \tag{10}$$

$$\geq \frac{\sqrt{\sum_{x} m_{x,T} \cdot \Delta(x,T)^{2}} + \sqrt{\sum_{x} m_{x,S} \cdot \Delta(x,S)^{2}}}{\sqrt{\beta}}$$
 (11)

$$\geq \frac{\sqrt{\beta_T \cdot \sum_x m_{x,T}^1 \cdot \Delta(x,T)^2} + \sqrt{\beta_S \sum_x m_{x,S}^1 \cdot \Delta(x,S)^2}}{\sqrt{\beta}}$$
 (12)

$$\geq \sqrt{\sum_{x} m_{x,T}^{1} \cdot \Delta(x,T)^{2}} + \sqrt{\sum_{x} m_{x,S}^{1} \cdot \Delta(x,S)^{2}}$$

$$\tag{13}$$

$$\geq \sum_{x} m_{x,T}^{1} \cdot \Delta(x,T) - \sum_{x} m_{x,S}^{1} \cdot \Delta(x,S)$$
(14)

$$= Imb(h^{\star}) - Imb(h) \tag{15}$$

where: (9) follows directly by applying f_1 to the expression we derived in Equation (8) for the loss difference; the transition in (10) follows from the fact that $\sqrt{2(a+b)} \ge \sqrt{a} + \sqrt{b}$ for $a,b \ge 0$; the transition in (11) follows by the definition of μ as min $\{\mu_T, \mu_S\}$; the transition in (12) uses the fact that $m_{x,g} \ge m_{x,g}^1 \cdot \beta_g$, which follows from the law of total probability:

$$m_{x,g} = \Pr_{D_g}(X = x)$$

$$= \sum_{y} \left[\Pr_{D_g}(X = x, Y = y) \right]$$

$$= \sum_{y} \left[\Pr_{D_g}(X = x | Y = y) \cdot \Pr_{D_g}(Y = y) \right]$$

$$\geq \Pr_{D_g}(X = x | Y = 1) \cdot \Pr_{D_g}(Y = 1)$$

$$= m_{x,g}^1 \cdot \beta_g$$

The transition in (13) follows by the definition of β as min $\{\beta_T, \beta_S\}$, and the transition in (14) is an application of Jensen's inequality (in the finite sum version), which states that $\varphi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_i}$ for concave φ . Indeed, using the fact that $\sum_x m_{x,g}^1 = 1$,

$$\sqrt{\sum_{x} m_{x,g}^{1} \cdot \Delta(x,g)^{2}} = \sqrt{\frac{\sum_{x} m_{x,g}^{1} \cdot \Delta(x,g)^{2}}{\sum_{x} m_{x,g}^{1}}} \ge \frac{\sum_{x} m_{x,g}^{1} \cdot \sqrt{\Delta(x,g)^{2}}}{\sum_{x} m_{x,g}^{1}} = \frac{\sum_{x} m_{x,g}^{1} \cdot |\Delta(x,g)|}{\sum_{x} m_{x,g}^{1}} = \sum_{x} m_{x,g}^{1} \cdot |\Delta(x,g)|$$

Finally, the transition in (15) uses the expression we derived for the imbalance difference in Equation (7).

D Proof of Theorem 2

The proof of the first part of Theorem 2 will be based on the following lemma, which we state and prove first.

Lemma D.1. Fix any $0 \le a \le b \le 1$. If $\gamma \ge b$, then $h_a = a \cdot h_1 + (1-a) \cdot h_0$ doesn't γ -disqualify $h_b = b \cdot h_1 + (1-b) \cdot h_0$.

Proof. We want to prove that h_a doesn't γ -disqualify h_b :

$$Imb(h_b) - Imb(h_a) \le f_{\gamma} \left(\min \left\{ 0, \ell(h_a) - \ell(h_b) \right\} \right)$$

By definition, h_a is less accurate and more balanced than h_b , so we can re-write the above as

$$Imb(h_b) - Imb(h_a) \le f_{\gamma} \left(\ell(h_a) - \ell(h_b) \right) \tag{16}$$

Let's start with the left-hand-side. Since imbalance is a linear operator, we have that in general, $Imb(h_{\alpha}) = Imb(\alpha \cdot h_1 + (1-\alpha) \cdot h_0) = \alpha \cdot Imb(h_1) + (1-\alpha) \cdot Imb(h_0) = \alpha \cdot Imb(h_1)$, where the last transition follows from the fact that h_0 is, by definition, perfectly balanced. We therefore have that

$$Imb(h_b) - Imb(h_a) = (b - a) \cdot Imb(h_1)$$

Next, recall that h_1 is 1—fair, so h_0 doesn't 1—disqualify it:

$$Imb(h_1) - Imb(h_0) \le f_1 \left(\min \left\{ 0, \ell(h_0) - \ell(h_1) \right\} \right)$$

Equivalently, $Imb(h_1) \le f_1(\ell(h_0) - \ell(h_1))$. Putting it together, for the inequality in (16) to hold, it suffices that

$$(b-a) \cdot f_1 (\ell(h_0) - \ell(h_1)) \le f_{\gamma} (\ell(h_a) - \ell(h_b)) \tag{17}$$

Denote $c = \ell(h_0) - \ell(h_1)$. We will be using the following fact:

$$\ell(h_a) - \ell(h_b) = (b - a) \cdot (2 - a - b) \cdot c \tag{18}$$

We will first show how, assuming Equation (18) is true, we can complete the proof of the lemma. Using the definition of f as $f_{\gamma}(a) = \sqrt{\frac{2\gamma a}{\mu\beta}}$, Equation (17) is equivalent to

$$(b-a)\cdot\sqrt{\frac{2\cdot c}{\mu\beta}}\leq\sqrt{\frac{2\cdot c\cdot \gamma\cdot (b-a)(2-a-b)}{\mu\beta}}$$

Or $(b-a) \le \gamma \cdot (2-a-b)$. It's therefore left to show that $\gamma \ge b$ implies

$$\gamma \ge \frac{b-a}{2-a-b} \tag{19}$$

Denote $b = 1 - \tau$ and $a = 1 - \tau - \varepsilon$; where the assumption $0 \le a \le b \le 1$ implies $\tau \in [0,1]$ and $\varepsilon \in [0,1-\tau]$. Indeed:

$$\frac{b-a}{2-a-b} = \frac{\varepsilon}{2\tau + \varepsilon}$$

$$= \frac{1}{1+2\tau/\varepsilon}$$

$$\leq \frac{1}{1+2\tau/(1-\tau)}$$

$$= \frac{1-\tau}{1+\tau}$$

$$\leq 1-\tau$$

$$= b$$

$$\leq \gamma$$

To conclude the lemma, we return to the proof of the fact in Equation (18):

$$\ell(h_a) - \ell(h_b) = (b-a) \cdot (2-a-b) \cdot \ell(h_0) - \ell(h_1)$$

Indeed, note that for a classifier h, we can write the squared loss as

$$\ell(h) = \underset{x,y}{\mathbf{E}} [\ell(h(x), y)]$$

$$= \underset{x}{\mathbf{E}} \left[h_1(x) \cdot (1 - h(x))^2 + (1 - h_1(x)) \cdot h(x)^2 \right]$$

$$= \underset{x}{\mathbf{E}} \left[h_1(x) - 2h(x)h_1(x) + h(x)^2 \right]$$

This means that the difference $\ell(h_a) - \ell(h_b)$ equals

$$\ell(h_a) - \ell(h_b) = \underset{x}{\mathbf{E}} \left[-2h_a(x)h_1(x) + h_a(x)^2 + 2h_b(x)h_1(x) - h_b(x)^2 \right]$$

$$= \underset{x}{\mathbf{E}} \left[(h_a(x) - h_b(x)) \cdot (h_a(x) + h_b(x) - 2h_1(x)) \right]$$
(20)

Additionally, we can write $h_a - h_b$ and $h_a + h_b$ in terms of h_1 and h_0 , as follows:

$$h_a - h_b = (a - b) \cdot (h_1 - h_0)$$

 $h_a + h_b = (a + b) \cdot h_1 - (a + b - 2) \cdot h_0$

So $h_a + h_b - 2h_1 = (a + b - 2) \cdot (h_1 - h_0)$, and plugging this back into the expression in (20), we get:

$$\ell(h_a) - \ell(h_b) = (a - b) \cdot (a + b - 2) \cdot \underset{x}{\mathbf{E}} \left[(h_1(x) - h_0(x))^2 \right]$$

In particular, for a = 0, b = 1 we get:

$$\ell(h_0) - \ell(h_1) = \mathbf{E}_{\mathbf{x}} \left[(h_1(\mathbf{x}) - h_0(\mathbf{x}))^2 \right]$$

which is exactly what we wanted to show.

We will now use the lemma to prove Theorem 2.

Proof. (i). Let $h' = \alpha \cdot h_1 + (1 - \alpha) \cdot h'_0$ be some classifier in $\tilde{\mathcal{H}}$, where h'_0 is some 0-fair classifier and $0 \le \alpha \le 1$. We need to prove that h' does not γ -disqualify h_{γ} . To do so, we argue that (a) h_{α} doesn't γ -disqualify h_{γ} and (b) if h_{α} doesn't γ -disqualify h_{γ} , then neither does h'.

For (a), note that this is an immediate corollary from Lemma D.1. First, for h' to disqualify h_{γ} it must be more balanced, so $\alpha < \gamma$. But then the lemma guarantees that h_{α} doesn't disqualify h_{γ} (with h_{α} in the role of h_a and h_{γ} in the role of h_b , so clearly the assumption of the lemma is true because $\gamma \geq b$ is the same as $\gamma \geq \gamma$).

For (b), we will be using the fact that $Imb(h') = Imb(h_{\alpha})$ (which follows directly from the definition), our claim in (a), and the fact that $\ell(h_{\alpha}) \leq \ell(h')$ (which we will prove promptly), in order. This yields:

$$Imb(h_{\gamma}) - Imb(h') = Imb(h_{\gamma}) - Imb(h_{\alpha}) \le f_{\gamma}(\ell(h_{\alpha}) - \ell(h_{\gamma})) \le f_{\gamma}(\ell(h') - \ell(h_{\gamma}))$$

which means that h' doesn't γ -disqualify h_{γ} , which is what we wanted to show.

It is left to prove the fact that $\ell(h_{\alpha}) \leq \ell(h')$, or $\ell(\alpha \cdot h_1 + (1-\alpha) \cdot h'_0) \leq \ell(\alpha \cdot h_1 + (1-\alpha) \cdot h_0)$. First, note that $h' - h_{\alpha} = (1-\alpha) \cdot (h'_0 - h_0)$ and $h' + h_{\alpha} = 2\alpha \cdot h_1 + (1-\alpha) \cdot (h'_0 + h_0)$. The latter means that $h' + h_{\alpha} - 2h_1 = (1-\alpha) \cdot (h'_0 + h_0 - 2h_1)$. Now,

$$\ell(h') - \ell(h_{\alpha}) = \mathbf{E}[h_{1} - 2h'h_{1} + h'^{2} - h_{1} + 2h_{\alpha}h_{1} - h_{\alpha}^{2}]$$

$$= \mathbf{E}[-2h'_{\alpha}h_{1} + h'^{2} + 2h_{\alpha}h_{1} - h_{\alpha}^{2}]$$

$$= \mathbf{E}[2h_{1}(h_{\alpha} - h') + (h' - h_{\alpha})(h' + h)]$$

$$= \mathbf{E}[(h' - h_{\alpha})(h' + h_{\alpha} - 2h_{1})]$$

$$= \mathbf{E}[(1 - \alpha) \cdot (h'_{0} - h_{0})(1 - \alpha) \cdot (h'_{0} + h_{0} - 2h_{1})]$$

$$= (1 - \alpha)^{2} \cdot \mathbf{E}[(h'_{0} - h_{0})(h'_{0} + h_{0} - 2h_{1})]$$

$$= (1 - \alpha)^{2} \cdot \mathbf{E}[h'_{0}(x)^{2} - 2h_{1}h'_{0} - h_{0}(x)^{2} + 2h_{1}h_{0}]$$

$$= (1 - \alpha)^{2} \cdot \mathbf{E}[h'_{0}(x)^{2} - 2h_{1}h'_{0} - h_{0}(x)^{2} + 2h_{1}h_{0} + h_{1}^{2} - h_{1}^{2}]$$

$$= (1 - \alpha)^{2} \cdot \left[\mathbf{E}[h'_{0}(x)^{2} - 2h_{1}h'_{0} + h_{1}(x)^{2}] - \mathbf{E}[h_{0}(x)^{2} - 2h_{1}h_{0} + h_{1}(x)^{2}]\right]$$

$$= (1 - \alpha)^{2} \cdot \left[\ell(h'_{0}) - \ell(h_{0})\right]$$

$$= (1 - \alpha)^{2} \cdot \left[\ell(h'_{0}) - \ell(h_{0})\right]$$

$$> 0$$

The final transition uses the fact that $\alpha \le 1$ and $[\ell(h'_0) - \ell(h_0)] \ge 0$ (which is true because h_0 was defined to be the most accurate 0-fair predictor).

(ii). We turn to prove the second part of Theorem 2. Recall we want to show that there exists a 0-fair classifier $h_0' \neq h_0$, such that $h_\gamma' = \gamma \cdot h_1 + (1 - \gamma) \cdot h_0' \in \tilde{\mathcal{H}}$ is not $(\gamma, \tilde{\mathcal{H}})$ -fair.

We will construct such an instance, as follows. Let S^+, S^-, T^+, T^- denote the positive and negative subsets of S and T, respectively (Note that these groups may not be explicitly defined in \mathcal{X}). Suppose that the features \mathcal{X} are such that only the following subsets can be identified from \mathcal{X} : S_1 (consisting of all of S^+ and half of S^-), T_1 (consisting of all of T^+ and half of T^-) and T^+ . This means the optimal classifier is

$$h_1(x) = \begin{cases} 1, & x \in T^+ \\ 2/3, & x \in S_1 \\ 0, & \text{otherwise} \end{cases}$$

Note that h_1 has an imbalance of 1/3. Consider two 0-fair classifiers:

$$g_0(x) = \begin{cases} 2/3, & x \in T_1 \\ 2/3, & x \in S_1 \\ 0, & \text{otherwise} \end{cases}, \quad g'_0(x) = 0.5$$

And consider $h'_{\gamma} = \gamma \cdot h_1 + (1 - \gamma) \cdot g'_0$ vs $h_{\gamma} = \gamma \cdot h_1 + (1 - \gamma) \cdot g_0$. We want to use the second to γ -disqualify the first. Towards this, consider the classifier $h_{\gamma-\varepsilon}$ – a mix of h_1 and g_0 that puts slightly less weight on h_1 . Relative to h'_{γ} , this classifier *improves* the imbalance by ε . However, we argue that it is also more accurate (for a sufficiently small ε); indeed, as we take ε to zero, the difference $\ell(h_{\gamma-\varepsilon} - \ell(h'_{\gamma}))$ tends to the difference $\ell(h_{\gamma}) - \ell(h'_{\gamma})$, and the latter is strictly positive for every γ (since g_0 is more accurate than g'_0). We therefore have that for any level of γ , for a sufficiently small ε , $h_{\gamma-\varepsilon}$ is both more accurate and more balanced than h'_{γ} ; thus, by definition, $h_{\gamma-\varepsilon}$ γ -disqualifies h'_{γ} . Since $h_{\gamma-\varepsilon} \in \tilde{\mathcal{H}}$, we conclude that h'_{γ} is not $(\gamma, \tilde{\mathcal{H}})$ -fair, as required.

E Proof of Theorem 3

Before we turn to proving Theorem 3, we formalize the notion of "reasonable" scaling function.

Definition E.1. A scaling function $f : \mathbb{R}^+ \times [0,1] \to \mathbb{R}^+$ is legal if it can be written as

$$f(\gamma, a) = t_1(\gamma) \cdot t_2(a)$$

such that the following requirements hold: (i) t_1 and t_2 are non-decreasing, (ii) $t_1(\cdot)$ is zero at $\gamma=0$ and tends to infinity as $\gamma \to \infty$ (and only as $\gamma \to \infty$), and (iii) $t_2(0)=0$.

We will often write the scaling function as $f_{\gamma}(a)$ (i.e., as a function from [0,1] to \mathbb{R}^+ , parameterized by $\gamma \in \mathbb{R}^+$). Finally, we use \mathcal{F} to denote the set of all such legal scaling functions.

To prove Theorem 3, we again consider the simple example in which $\mathcal{X} = \emptyset$ (so the only information available for prediction is group membership). In this case, the Bayes optimal classifiers for the squared loss and expected zero one loss, respectively, are:

$$h_{\ell^2}^{\star}(x,g) = \beta_g, \qquad h_{\ell^{0-1}}^{\star}(x,g) = \mathbf{1}[\beta_g > 0.5]$$

Suppose that for some positive τ , $\beta_T = 0.5 + \tau$ and $\beta_S = 0.5 - \tau$. Note that in this case $h_{\ell^{0-1}}^{\star}$ is maximally imbalanced, since it predicts 1.0 for all members of T and 0.0 for all members of S. Now, consider $h_{\ell^2}^{\star}$ as the alternative classifier. When does $h_{\ell^2}^{\star}$ 1-disqualify $h_{\ell^{0-1}}^{\star}$? In terms of imbalance, its imbalance is $\beta_T - \beta_S = 2\tau$, so the improvement in imbalance is $1.0 - 2\tau$. In terms of accuracy, we have:

$$\ell(h_{\ell^{0-1}}^{\star}) = \mu_S \beta_S + \mu_T (1 - \beta_T)$$

$$\ell(h_{\ell^2}^{\star}) = 2 \cdot [\mu_S \beta_S \cdot (1 - \beta_S) + \mu_T \beta_T \cdot (1 - \beta_T)]$$

So the difference in loss is

$$\ell(h_{\ell^2}^{\star}) - \ell(h_{\ell^{0-1}}^{\star}) = \mu_S \beta_S (2 - 2\beta_S - 1) + \mu_T (1 - \beta_T) (2\beta_T - 1)$$

= $2\tau \cdot [\mu_S \beta_S + \mu_T (1 - \beta_T)]$

In order for $h_{\rho 0-1}^{\star}$ to be γ -fair, the following must hold:

$$1 - 2\tau \le f_{\gamma}(2\tau \cdot [\mu_S \beta_S + \mu_T(1 - \beta_T)]) = t_1(\gamma) \cdot t_2(2\tau \cdot [\mu_S \beta_S + \mu_T(1 - \beta_T)])$$
 (22)

Note that if f is a legal scaling function, $t_1(\gamma)$ is bounded for $\gamma < \infty$, e.g. by M. This means that as we take $\tau \to 0$, the RHS of Equation (22) approaches $M \cdot 0 = 0$, whereas the LHS approaches 1.0. This shows that there cannot be a single $f \in \mathcal{F}$ that guarantees $h_{\ell^0-1}^{\star}$ satisfies γ – Pareto Fairness for $\gamma < \infty$.

F Proof of Theorem 4

F.1 Existence of (γ, \mathcal{H}) -fair classifiers

In general, it's not always the case that \mathcal{H} itself contains a classifier that is γ -fair w.r.t \mathcal{H} . This is the case, however, for $\Delta(\mathcal{H})$, the class of convex combinations of classifiers from \mathcal{H} .

Lemma F.1. For every compact $\mathcal{H} \neq \phi$ and $\gamma \in [0, \infty)$, the class $\Delta(\mathcal{H})$ always contains a (γ, \mathcal{H}) -fair classifier.

Proof. Let h be some classifier in \mathcal{H} . If Imb(h)=0, then h is perfectly balanced and in particular (γ,\mathcal{H}) -fair. Otherwise assume w.l.o.g that the imbalance is in favor of T. Next, let h' be the classifier that minimizes the directed loss imbalance (in the direction $T\to S$) in \mathcal{H} . Now, there are exactly two cases: (i) $dImb(h')\geq 0$, (ii) dImb(h')<0. In the first case, we claim that h' itself is (γ,\mathcal{H}) -fair. Indeed, according to our definition, for some other $h''\in\mathcal{H}$ to disqualify h', h'' must strictly improve the directed imbalance – but by definition no classifier in \mathcal{H} can do that. In the second case, we have that dImb(h)>0 and dImb(h')<0. From the linearity of the directed imbalance, there must exist a such that $dImb(a\cdot h+(1-a)\cdot h')=0$. This is a classifier in $\Delta(\mathcal{H})$ that is perfectly balanced and therefore also (γ,\mathcal{H}) -fair, which concludes the proof.

⁷The assumption that \mathcal{H} is compact guarantees this argmin exists. For example, any finite \mathcal{H} , or an infinite class parametrized by a compact space such as \mathcal{H}_w (linear or logistic classifiers with a bounded norm).

F.1.1 Approximate Fairness

For a quantity P that depends on the distribution \mathbb{P} , we use \widehat{P} to denote the empirical counterpart of P as calculated w.r.t a fixed sample $D \sim \mathbb{P}$. Note that our definition of disqualification is not immediately applicable when working with finite samples: for example, even a perfectly balanced classifier, Imb(h)=0, will have some level of imbalance on a random sample: $\widehat{Imb}(h)\neq 0$. The same is true with respect to the loss; e.g., even the most accurate classifier (which should always be ∞ -fair) might not have the optimal loss on a random sample and may therefore be disqualified. Thus, to guarantee generalization we define the following approximate variants of γ -disqualification and γ -fairness.

Definition F.2 (Approximate disqualification). A classifier h' is said to $(\alpha_1, \alpha_2, \gamma)$ -disqualify h if

$$dLossImb(h; \ell_B) - dLossImb(h'; \ell_B) > \alpha_1 + f_{\gamma} \left(\max \left\{ 0, \ \ell_A(h') - \ell_A(h) + \alpha_2 \right\} \right)$$
 (23)

Definition F.3 (Approximate fairness). We say that a classifier h is $(\alpha_1, \alpha_2, \gamma)$ -fair w.r.t \mathcal{H} if no classifier in \mathcal{H} $(\alpha_1, \alpha_2, \gamma)$ -disqualifies it.

For simplicity, we will sometimes simply say that h is (α, γ) -fair w.r.t \mathcal{H} , where $\alpha = (\alpha_1, \alpha_2)$ are the approximation parameters (for imbalance and loss, respectively). Additionally, when Definitions (F.2) and (F.3) are computed w.r.t a finite sample $D \sim \mathbb{P}^m$ (as opposed to w.r.t \mathbb{P}) itself, then we say that h' empirically disqualifies h and that h is empirically fair, respectively.

With the definitions of approximate and empirical fairness in place, we define a notion of uniform convergence of a class \mathcal{H} w.r.t our notion of fairness, as follows.

Definition F.4 (Uniform convergence w.r.t fairness). We say that a class \mathcal{H} has the uniform convergence property with respect to approximate-fairness with sample complexity $m_{\epsilon_1,\epsilon_2,\delta}^{Pareto}$ if for any distribution \mathbb{P} , w.p $1-\delta$ over the choice of $D\sim\mathbb{P}^m$ for $m\geq m_{\epsilon_1,\epsilon_2,\delta}^{\gamma-Pareto}$, the following holds simultaneously for every $h\in\mathcal{H}$: there exist $\alpha=(\alpha_1,\alpha_2)$ and $\hat{\alpha}=(\hat{\alpha}_1,\hat{\alpha}_2)$ such that: (i) h is (α,γ) -fair on the underlying distribution \mathbb{P} , (ii) h is empirically $(\hat{\alpha},\gamma)$ -fair on D; (iii) $|\alpha_i-\hat{\alpha}_i|\leq \varepsilon_i$ for i=1,2.

Importantly, we show that standard uniform convergence for loss and imbalance indeed guarantees uniform convergence w.r.t our notion of fairness. This is important, since it will allow us to solve the *fair risk minimization* on a finite sample, and argue that the results transfer (approximately) to the underlying distribution.

Lemma F.5. If \mathcal{H} has the uniform convergence property w.r.t loss with sample complexity $m_{\varepsilon,\delta}^{\text{Loss}}$ and w.r.t imbalance with sample complexity $m_{\varepsilon,\delta}^{\text{Imb}}$, then it has uniform convergence property w.r.t approximate fairness, with sample complexity $m_{\varepsilon_1,\varepsilon_2,\delta}^{\text{Pareto}} = \max\{m_{\varepsilon_1/2,\delta/2}^{\text{Imb}}, m_{\varepsilon_2/2,\delta/2}^{\text{Loss}}\}$.

Proof. Suppose $m \ge \max\{m^{Imb}_{\varepsilon_1/2,\delta/2}, m^{Loss}_{\varepsilon_2/2,\delta/2}\}$ and consider some classifier $h \in \mathcal{H}$. Suppose that it is (α, γ) -fair w.r.t \mathcal{H} on $D \sim \mathbb{P}^m$. Let h' be any other classifier in \mathcal{H} . Using the uniform convergence assumptions, we have that w.p at least $1 - \delta$

$$\begin{split} Imb(h) - Imb(h') &\leq 2 \cdot \varepsilon_1 / 2 + \widehat{Imb}(h) - \widehat{Imb}(h') \\ &\leq \varepsilon_1 + \alpha_1 + c \cdot \sqrt{\gamma \cdot \left[\widehat{\ell}(h') - \widehat{\ell}(h) + \alpha_2\right]} \\ &= \varepsilon_1 + \alpha_1 + c \cdot \sqrt{\gamma \cdot \left[\ell(h') - \ell(h) + \alpha_2 + \varepsilon_2\right]} \end{split}$$

So h is (α', γ) -fair w.r.t \mathcal{H} on \mathbb{P} , where $\alpha'_i = \alpha_i + \varepsilon_i$, which implies the required.

F.1.2 Approximate Pareto Frontier of \mathcal{H}

The Empirical Pareto frontier of a class of classifiers \mathcal{C} (w.r.t loss and imbalance in a fixed direction, as measured on a sample D) is the collection of all Pareto-efficient classifiers in \mathcal{C} ; that is, thinking of a classifier h as a two-dimensional point $h=(\varphi,\tau)\in[0,1]\times[-1,1]$ where $\widehat{dImb}(h)=\tau$ and $\widehat{\ell}(h)=\varphi$, these are all the points in \mathcal{C} that correspond to classifiers that achieve the best possible loss (of any classifier in \mathcal{C}) without exceeding a specific level of imbalance. We define an ε -approximation to the Empirical Pareto frontier as follows:

$$\widehat{\mathsf{PF}}(\varepsilon, \mathcal{C}; D) \qquad \{h(\tau) \mid \tau = -1, -1 + \varepsilon, \dots, 1 - \varepsilon, 1\}$$
 (24)

where

$$h(\tau) \in \arg\min_{h \in \mathcal{C}} \widehat{\ell}(h) \quad \text{subject to} \quad \widehat{Imb}(h) \le \tau$$
 (25)

That is, $\widehat{\mathsf{PF}}(\varepsilon,\mathcal{C})$ is a collection of at most $2/\varepsilon$ classifiers from \mathcal{C} , each of which is optimal for their maximal imbalance level. We refer to $\widehat{\mathsf{PF}}(\varepsilon=0,\mathcal{C})$ as the *full* Empirical Pareto frontier.

Disqualifying on the Pareto frontier. Given α and γ , we say that a classifier is (empirically) disqualified on the ε -Pareto-frontier of $\mathcal C$ if there is a classifier in $\widehat{\mathsf{PF}}(\varepsilon,\mathcal C)$ that (α,γ) (empirically) disqualifies it. Note that by definition, if h is a classifier which no classifier on the *full* Pareto frontier disqualifies, then h is (empirically) fair w.r.t $\mathcal C$. When we use $\widehat{\mathsf{PF}}(\varepsilon,\mathcal C)$, however, we incur an additional factor of ε in the additive imbalance slack:

Lemma F.6. Fix a classifier p, parameters α , γ and a class C. If p is not $(\alpha_1, \alpha_2, \gamma)$ -empirically-disqualified on $\widehat{\mathsf{PF}}(\varepsilon, C)$, then it is not $(\alpha_1 + \varepsilon, \alpha_2, \gamma)$ -empirically-disqualified on $\widehat{\mathsf{PF}}(\varepsilon = 0, C)$, the full Pareto frontier.

Proof. Let h be any classifier on the full Pareto frontier, with empirical imbalance τ . Denote by $\tau' \geq \tau$ the closest imbalance on the ε-Pareto frontier, and h' the optimal classifier at this imbalance level. Note that by the definition of $\widehat{\mathsf{PF}}(\varepsilon, \Delta(H))$, $\tau \leq \tau' \leq \tau + \varepsilon$ and $\widehat{\ell}(h') \leq \widehat{\ell}(h)$. Now, we can use the fact that h' is on the ε-Pareto frontier and therefore doesn't empirically-disqualify p:

$$\begin{split} \widehat{Imb}(p) - \widehat{Imb}(h) &= \widehat{Imb}(p) - \tau \\ &\leq \widehat{Imb}(p) - (\tau' - \varepsilon) \\ &= \varepsilon + \widehat{Imb}(p) - \widehat{Imb}(h') \\ &\leq \varepsilon + \alpha_1 + f_{\gamma}(\widehat{\ell}(h') - \widehat{\ell}(p) + \alpha_2) \\ &\leq \varepsilon + \alpha_1 + f_{\gamma}(\widehat{\ell}(h) - \widehat{\ell}(p) + \alpha_2) \end{split}$$

From which we conclude that *h* doesn't empirically $(\alpha_1 + \varepsilon, \alpha_2, \gamma)$ -disqualify *p*, as required.

We will also use the fact that if there is a classifier that is empirically fair, then there is also one that is similarly empirically fair *and* is on the approximate Pareto frontier.

Lemma F.7. *If* p *is empirically* $(\alpha_1, \alpha_2, \gamma)$ -fair w.r.t \mathcal{H} , then there is a classifier p' on the ε -Pareto frontier oh \mathcal{H} which is $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -fair w.r.t \mathcal{H} .

Proof. Denote $\tau = \widehat{Imb}(p)$. Let $\tau' \geq \tau$ denote the closest imbalance level that corresponds to an imbalance on the ε -Pareto frontier, and denote $p' = h(\tau') \in \widehat{\mathsf{PF}}(\varepsilon, \Delta(\mathcal{H}))$ the classifier that is optimal at this level of imbalance. Now, let h be any other classifier - we'll prove it doesn't $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -empirically-disqualify p':

$$\begin{split} \widehat{Imb}(p') - \widehat{Imb}(h) &\leq \tau' - \widehat{Imb}(h) \\ &\leq \tau + \varepsilon - \widehat{Imb}(h) \\ &= \varepsilon + \widehat{Imb}(p) - \widehat{Imb}(h) \\ &\leq \varepsilon + \alpha_1 + f_{\gamma}(\widehat{\ell}(h) - \widehat{\ell}(p) + \alpha_2) \\ &\leq \varepsilon + \alpha_1 + f_{\gamma}(\widehat{\ell}(h) - \widehat{\ell}(p') + \alpha_2) \end{split}$$

Computing the approximate Pareto Frontier. We will be working with the approximate Pareto frontier for $\Delta(\mathcal{H})$, the class of convex combinations of classifiers in \mathcal{H} . In some simple cases, $\widehat{\mathsf{PF}}(\varepsilon,\Delta(\mathcal{H}))$ can be computed efficiently. For example, when $\mathcal{H}=\mathcal{H}_w$ is the class of linear classifiers over \mathbb{R}^d with bounded norm (in which case $\Delta(\mathcal{H})=\mathcal{H}$), $\varphi(\tau)$ can be obtained as the solution to a convex program with d variables (since the objective is convex in w, and the imbalance constraints are linear in w). Therefore in this case $\widehat{\mathsf{PF}}(\varepsilon,\Delta(\mathcal{H}))$ can be computed in time $\mathsf{poly}(1/\varepsilon,d,|D|)$.

F.2 Learning optimal fair classifiers using the Pareto Frontier

A natural objective is to find the *optimal* fair classifier in $\Delta(\mathcal{H})$ (which always contains a γ -fair classifier). That is, given a class \mathcal{H} parameters α, γ and a sample D, find the optimal $h \in \Delta(\mathcal{H})$ that is empirically (α, γ) -fair w.r.t \mathcal{H} . We refer to this as the Fair-ERM problem:

FairERM
$$(\alpha, \gamma, \mathcal{H})$$
 $\min_{h \in \Delta(\mathcal{H})} \widehat{\ell}(h)$ s.t h is empirically (α, γ) – fair w.r.t \mathcal{H} (26)

The next proposition proves that an approximation to the Pareto frontier can be used to efficiently obtain an approximation to the FairERM problem.

Proposition F.8. Fix a dataset D, a class \mathcal{H} and parameters $\alpha_1, \alpha_2, \gamma$. Then, for every $\varepsilon \leq \alpha_1$ the following is true: Given oracle access to $\widehat{\mathsf{PF}}(\varepsilon, \Delta(H))$, there is an efficient algorithm A for finding a classifier h(A) such that: (1) h(A) is empirically $(\alpha_1, \alpha_2, \gamma)$ -fair $w.r.t \, \mathcal{H}$, and (2) $\widehat{\ell}(h(A)) \leq \widehat{\ell}(h)$, where h is any classifier that is empirically $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -fair $w.r.t \, \mathcal{H}$.

Note that by using the approximation to the Pareto frontier we incur a degradation in the accuracy guarantee: we are only outputting a classifier whose accuracy is competitive with the optimal

 $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -fair classifier (as opposed to with the optimal $(\alpha_1, \alpha_2, \gamma)$ -fair classifier). Without assuming anything about \mathcal{H} , this accuracy gap could be substantial: in principle, \mathcal{H} could contain two classifiers which differ only minimally in imbalance but significantly in accuracy. This highlights that the strength of the guarantee is related to the *Lipschitzness* of the function $\varphi(\tau; \Delta(\mathcal{H}))$ that returns the optimal loss in $\Delta(\mathcal{H})$ at a given level of imbalance (see Equation 25).

```
Initialize FairClassifiers = [ ]
for h \in \widehat{\mathsf{PF}}(\varepsilon, \Delta(H)) do
    fair = True
    for h' \in \widehat{\mathsf{PF}}(\varepsilon, \Delta(H)) do
         if h' empirically (\alpha_1 - \varepsilon, \alpha_2, \gamma)-disqualifies h then
              fair = False
              break
         end if
    end for
    if fair then
         FairClassifiers.append(h)
    end if
end for
if FairClassifiers == \emptyset: then
 ∣ return ⊥
end if
h(A) \leftarrow the most accurate classifier in FairClassifiers
```

Figure 2: ApproxFairERM $(\varepsilon, \alpha_1, \alpha_2, \gamma, \mathcal{H})$ returns the most accurate classifier in the set $\widehat{\mathsf{PF}}(\varepsilon, \Delta(H))$ that is not empirically $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -disqualified by another classifier in this set.

Proof. Consider the procedure $A \triangleq \mathsf{ApproxFairERM}(\varepsilon,\alpha_1,\alpha_2,\gamma,\mathcal{H})$ defined in Figure 2, which runs in time $O(1/\varepsilon^2)$. First, we argue that if the output of A is $h \neq \bot$, then h is $(\alpha_1,\alpha_2,\gamma)$ -empirically-fair. Note that ApproxFairERM only returns a classifier which no other classifier on the ε -Pareto frontier $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ disqualifies. Therefore, Lemma F.6 guarantees that no classifier on the full Pareto frontier $(\alpha_1,\alpha_2,\gamma)$ disqualifies it; this, in turn, guarantees the classifier is $(\alpha_1,\alpha_2,\gamma)$ -fair. Second, we argue that A always returns $h \neq \bot$. In light of the above, it's sufficient to argue that there always exists a classifier on the ε -Pareto-frontier that is $(\alpha_1,\alpha_2,\gamma)$ -fair. This follows by combining Lemma F.7 with the fact that there exists a classifier (not necessarily on the approximate Pareto frontier) that is $(\alpha_1 + \varepsilon, \alpha_2, \gamma)$ -fair.

It's left to prove that $\ell(h(A)) \leq \ell(h^*)$, where h^* is the optimal $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -empirically-fair classifier.

Let h denote the closest classifier (from above) to h^{\star} on $\mathsf{PF}(\varepsilon, \Delta(H))$, so that $\widehat{\ell}(h) \leq \widehat{\ell}(h^{\star})$ and $\widehat{\mathit{Imb}}(h) \leq \widehat{\mathit{Imb}}(h^{\star}) + \varepsilon$. We claim that h is empirically $(\alpha_1, \alpha_2, \gamma)$ -fair. To see this, let h' be some other classifier. Using the fact that h^{\star} is $(\alpha_1 - \varepsilon, \alpha_2, \gamma)$ -fair, we have:

$$\begin{split} \widehat{Imb}(h) - \widehat{Imb}(h') &\leq \varepsilon + \widehat{Imb}(h^{\star}) - \widehat{Imb}(h') \\ &\leq \varepsilon + \alpha_1 - \varepsilon + f_{\gamma}(\widehat{\ell}(h') - \widehat{\ell}(h^{\star}) + \alpha_2) \\ &\leq \alpha_1 + f_{\gamma}(\widehat{\ell}(h') - \widehat{\ell}(h^{\star}) + \alpha_2) \\ &\leq \alpha_1 + f_{\gamma}(\widehat{\ell}(h') - \widehat{\ell}(h) + \alpha_2) \end{split}$$

So h is indeed $(\alpha_1, \alpha_2, \gamma)$ -fair. Next, we note that by definition, h(A) is the optimal classifier on the ε -Pareto-frontier that is $(\alpha_1, \alpha_2, \gamma)$ -fair (this follows by the definition of ApproxFairERM, and (i)). Since h is, by construction, also on the ε -Pareto frontier, the fact it is $(\alpha_1, \alpha_2, \gamma)$ fair thus implies that $\widehat{\ell}(h(A)) \leq \widehat{\ell}(h)$. Since by definition $\widehat{\ell}(h) \leq \widehat{\ell}(h^*)$, we have that $\widehat{\ell}(h(A)) \leq \widehat{\ell}(h) \leq \widehat{\ell}(h^*)$, as required.