# Consider the Alternatives: Navigating Fairness-Accuracy Tradeoffs via Disqualification

**Guy N. Rothblum & Gal Yona, Weizmann Institute**

Group fairness notions can be in **direct conflict with accuracy** (even when the learner is well-intentioned).

We present a theoretical framework for reasoning about such tradeoffs in supervised ML. Our formalization draws inspiration from the disparate impact doctrine [1]:

"Disparate impact is not concerned with the intent or motive for a policy; where it applies, … first asks if 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*"

**Setup**: Predictors are mapping from features $X$ and group membership $A$ (specifying S, T) into [0,1].

**Definition** (Loss imbalance): The loss imbalance of $h$ (for $y=1$) in $T \rightarrow S$ is

$$dLossImb(h;\ell_B) = E_{x,y\sim S^+}[\ell_B(h(x),y)] - E_{x,y\sim T^+}[\ell_B(h(x),y)]$$

When $\ell_B$ is the exp. 0/1 loss, recovers *balance* [3] and *Equal Opportunity* [2].

## Disqualification

A classifier should be ***disqualified*** if there exists a fairer alternative that does not degrade accuracy by "too much".

"Too much" is quantified using a parameter γ≥0:

1 unit accuracy ≡ γ units fairness
1 unit fairness ≡ 1/γ units accuracy

Subtle point: requires specifying an appropriate *normalization* to bring fairness and accuracy to the same units.

**Definition** (γ-disqualification): A classifier h' γ-disqualifies h w.r.t $\ell_A, \ell_B$ if

$$dLossImb(h;\ell_B) - dLossImb(h';\ell_B) > f_\gamma([\ell_A(h') - \ell_A(h)]_+)$$

where loss imbalance is computed in the direction that $dLossImb(h;\ell_B) > 0$.

**Definition** (γ-fairness): h is (γ, H)-fair if it is not disqualified by any h' in H. In the unconstrained case, we say h is γ-fair.

## Scaling

How to instantiate the scaling function? Our approach: Consider the minimal level γ for which *the Bayes optimal predictor $h^*$ is not* γ *-disqualified by any other classifier*, and attempt to select the scaling function in a way that "anchors" the value at γ =1.

A (minimal) desirable property: guarantees the fairness of $h^*$ is invariant to scalar multiplications of $\ell_A$ (does not meaningfully change the fairness-accuracy trade-offs).

**Case studies**

1: Measuring accuracy using **squared loss**

We show a natural scaling function that satisfies the above requirement:

$$f_\gamma(a) = \sqrt{\gamma \cdot \frac{2a}{\min \eta_{S^+}, \eta_{T^+}}}$$

2: Measuring accuracy using **0/1 loss**

We show no "reasonable" function can satisfy the requirement. Intuitively, 0/1 loss is highly "Non-Lipschitz" w.r.t tradeoffs: tiny accuracy improvements can result in unbounded degradation in fairness.

**ERM subject to disqualification:**

We present an algorithm that, given a dataset D and parameter γ, finds an approximately optimal (γ,H)−fair classifier. The algorithm is stated as a reduction to the well studied task of approximating the Pareto frontier of H.

## Applications

**(1) Selection with $\gamma$:**

Example: Suppose we apply two strategies (fairness aware & unaware) on the Adult Income dataset, yielding **h** with accuracy (squared error) 0.149 but unfairness 0.11, and **h'** with accuracy 0.150 but improved unfairness of 0.0051. Disqualification tells us **we should prefer h' over h if fairness is 20x as important as accuracy is.**

(2) **Comparing strategies without γ:**

For a classifier h, compute the "effective unfairness" of h, $\hat{\gamma}(h)$, as the minimal value $\gamma$ for which there is another classifier that $\gamma$-disqualifies h.

[1] Big data's disparate impact; S Barocas, AD Selbst - Calif. L. Rev., 2016
[2] Equality of **opportunity** in supervised learning; M Hardt, E Price, N Srebro
[3] Inherent trade-offs in the fair determination of risk scores
J Kleinberg, S Mullainathan, M Raghavan