

---

# Fairness under Competition

---

Anonymous Author(s)

Affiliation

Address

email

## Appendix

### A Extensions

In this section we consider two extensions of the simplified model presented in Section 2. First, in Section A.1, we study more general utility functions of the borrowers. And second, in Section A.2, we study competition between more than two classifiers.

#### A.1 Generalization: different utility functions

In this section we restrict again to two classifiers, but return to the more general form of the utility function, namely,  $v : X \times A \times Y \times \mathbb{N} \mapsto \mathbb{R}$ , where  $\mathbb{N}$  is the set of non-negative integers. That is,  $v(x, a, y, r)$  is the utility of a borrower  $(x, a, y)$  when he receives  $r$  loan offers. However, we now drop the assumption that  $v(x, a, y, 1) = v(x, a, y, 2)$ . Instead, we assume that  $v(x, a, 1, 2) = k \cdot v(x, a, 1, 1)$  for some  $k \geq 1$ . As before, we assume that  $v(x, a, y, 0) = 0$ , and normalize  $v(x, a, y, 1) = 1$ . Our simplifying assumption in the main body of the paper was that  $k = 1$ , but now we consider all values of  $k \geq 1$ . We first generalize Proposition 1 to this setting, and then apply it to Example 1.

**Proposition 3** *For two EO classifiers  $c_1$  and  $c_2$  with false-negative rates  $\beta_1$  and  $\beta_2$ , the level of EOC is  $\sigma_1 \sigma_2 |(k-2)(\rho^0 - \rho^1)|$ . In the worst case, the level of EOC is  $|k-2| \cdot (\min\{\beta_1, \beta_2\} - \max\{0, \beta_1 + \beta_2 - 1\})$ .*

There are a few interesting observations to note. First, in all cases except  $k = 2$ , the level of EOC is positive, even though both classifiers are EO. Hence, our result is robust. Interestingly, however, there is a qualitative difference between the case  $k \in [1, 2)$  and  $k > 2$ , in that the group with lower utility under competition is different in the two parameter intervals. This last point is further explained at the end of this sub-section, in the context of the example.

**Proof of Proposition 3** Recall from the proof of Proposition 1 that

$$\Pr[B_1^a = B_2^a = 1] = \rho^a \sigma_1 \sigma_2 + (1 - \beta_1)(1 - \beta_2)$$

and

$$\Pr[B_1^a = B_2^a = 0] = \rho^a \sigma_1 \sigma_2 + \beta_1 \beta_2.$$

Thus, we have that

$$\begin{aligned} \Pr[B_1^a = 1 \cap B_2^a = 0] &= \Pr[B_1^a = 1] - \Pr[B_1^a = B_2^a = 1] \\ &= 1 - \beta_1 - \rho^a \sigma_1 \sigma_2 - (1 - \beta_1)(1 - \beta_2) \\ &= \beta_2(1 - \beta_1) - \rho^a \sigma_1 \sigma_2 \end{aligned}$$

and, analogously,

$$\Pr[B_1^a = 0 \cap B_2^a = 1] = \beta_1(1 - \beta_2) - \rho^a \sigma_1 \sigma_2.$$

27 Thus, for each  $a$ ,

$$\begin{aligned}
\mathbb{E}[d(X, A)|Y = 1, A = a] &= k \cdot \Pr[c_1(X, A) = c_2(X, A) = 1|Y = 1, A = a] \\
&\quad + \Pr[c_1(X, A) \neq c_2(X, A)|Y = 1, A = a] \\
&= k \cdot \Pr[B_1^a = B_2^a = 1] + \Pr[B_1^a \neq B_2^a] \\
&= k\rho^a\sigma_1\sigma_2 + k(1 - \beta_1)(1 - \beta_2) + \beta_2(1 - \beta_1) + \beta_1(1 - \beta_2) - 2\rho^a\sigma_1\sigma_2 \\
&= (k - 2)\rho^a\sigma_1\sigma_2 + k - (k - 1)(\beta_1 + \beta_2 - \beta_1\beta_2).
\end{aligned}$$

28 The level of EOC is thus

$$\begin{aligned}
&|\mathbb{E}[d(X, A)|Y = 1, A = 0] - \mathbb{E}[d(X, A)|Y = 1, A = 1]| \\
&= |(k - 2)\rho^0\sigma_1\sigma_2 - (k - 2)\rho^1\sigma_1\sigma_2| \\
&= \sigma_1\sigma_2 |(k - 2)(\rho^0 - \rho^1)|.
\end{aligned}$$

29 As in the proof of Proposition 1, the worst case scenario maximizes  $|\rho^0 - \rho^1|$ , and so, for example,  
30 maximizes  $\rho^0$  while minimizing  $\rho^1$  (or vice versa). Again, as before, the worst-case difference is  
31  $\frac{1}{\sigma_1\sigma_2}(\min\{\beta_1, \beta_2\} - \max\{0, \beta_1 + \beta_2 - 1\})$ , which leads to the claimed worst case level of EOC. ■

32 Proposition 3 yields the following simple corollary.

33 **Corollary 3** *If  $\beta = \beta_1 = \beta_2$  and  $\beta \leq 1/2$ , then the worst-case level of EOC is  $\beta \cdot |k - 2|$ .*

34 We can also use Proposition 3 to further analyze Example 1.

35 **Example 1 continued** Suppose again that all classifiers in the example—the third-party classifier,  
36 as well as the two firms’ individual classifiers—have the same false-negative rate  $\beta$ . Thus, a firm  
37 that uses its own classifier on  $A = 1$  and the third-party classifier on  $A = 0$  is in practice using an  
38 EO classifier. Suppose also that the firms’ individual classifiers  $c_1$  and  $c_2$  are uncorrelated on  $A = 1$ ,  
39 namely, that  $\rho^1 = 0$ . Finally, note that, since both firms use the same classifier on  $A = 0$  we also  
40 have  $\rho^0 = 1$ . Thus, even though each firm uses an EO classifier, the level of EOC is  $\beta(1 - \beta) \cdot k - 2$ .  
41 Furthermore, the utility of borrowers in group  $A = 0$  is  $k(1 - \beta)$ , since all borrowers who receive an  
42 offer actually receive two offers. In group  $A = 1$ , on the other hand, the expected utility of borrowers  
43 is  $k(1 - \beta)^2 + 2\beta(1 - \beta)$ .

44 Now,

$$k(1 - \beta)^2 + 2\beta(1 - \beta) - k(1 - \beta) = \beta(1 - \beta)(2 - k).$$

45 This implies that the level of EOC is  $\beta(1 - \beta)|2 - k|$ . However, note that, if  $k \in [1, 2)$  then the  
46 difference is positive, whereas if  $k > 2$  then the difference is negative. This implies that, in the  
47 former case, the expected utility is higher in group  $A = 1$ , whereas in the latter case, the expected  
48 utility is higher in group  $A = 0$ .

49 The main insights from this section are, first, that two EO classifiers can lead to a positive EOC even  
50 for more general utility functions, and second, that whether or not the “disadvantaged” group is the  
51 one that suffers under a positive EOC depends on the utility function (i.e., whether  $k \in [1, 2)$  or  
52  $k > 2$ ).

## 53 A.2 Generalization: more than two classifiers

54 Suppose that instead of two classifiers, the set  $L$  contains  $n$  classifiers. We provide two results.  
55 First, we characterize the worst-case EOC with  $n$  classifiers. Second, we generalize the analysis of  
56 Example 1 from Section 3 and show that, when classifiers are uncorrelated on one group but fully  
57 correlated on the other group, the EOC strictly increases with  $n$ .

58 Suppose there are  $n$  classifiers,  $c_1, \dots, c_n$ ,

59 **Proposition 4** *For  $n$  EO classifiers  $c_1, \dots, c_n$  with false-negative rates  $\beta_1, \dots, \beta_n$ , the worst-case  
60 level of EOC is  $\min_{i \in L} \beta_i - \max\{0, \sum_{j \in L} \beta_j - 1\}$ . If  $\beta_i = \beta \leq 1 - 1/n$  for all  $i$ , then the  
61 worst-case level of EOC is  $\beta$ .*

**Proof:** The worst case occurs when the probability that all classifiers misclassify positive instances is minimal in group  $A = 1$  and maximal in group  $A = 0$  (or vice versa). To maximize  $prc_1(X, A) = \dots = c_n(X, A) = 0|A = 0, Y = 1$  the overlap in instances on which each  $c_i$  misclassifies has to be maximal, which occurs when the instances are contained in one another. Formally, if  $\beta_1 \leq \dots \leq \beta_n$ , then

$$c_n(x, 0) = 1 \Rightarrow \dots \Rightarrow c_1(x, 0) = 1.$$

In this case,

$$prc_1(X, A) = \dots = c_n(X, A) = 0|A = 0, Y = 1 = \beta_n = \min_{i \in L} \beta_i.$$

To minimize  $prc_1(X, A) = \dots = c_n(X, A) = 0|A = 1, Y = 1$  the overlap in instances on which each  $c_i$  misclassifies has to be minimal, which occurs when the instances are maximally disjoint. Equivalently, the overlap in positive instances in which the  $c_i$ 's correctly classify is maximally disjoint. If  $\sum_j (1 - \beta_j) \geq 1$  then

$$\{(x, 1, 1) : c_1(x, 1) = \dots = c_n(x, 1) = 0\} = \emptyset.$$

Otherwise, in the worst case each classifier  $c_i$  correctly classifies a unique set of  $1 - \beta_i$  positive instances, and in this case

$$\Pr[c_1(X, A) = \dots = c_n(X, A) = 0|A = 0, Y = 1] = \sum_{j \in L} \beta_j - 1.$$

Thus, the worst-case level of EOC is  $\min_{i \in L} \beta_i - \max\{0, \sum_{j \in L} \beta_j - 1\}$ .

Finally, if  $\beta_i = \beta \leq 1 - 1/n$  for all  $i$  then  $\min_{i \in L} \beta_i = \beta$  and  $\max\{0, \sum_{j \in L} \beta_j - 1\} = 0$ . ■

We now analyze a further extension of Example 1, in which we show that the level of EOC is strictly increasing in the number of lenders.

**Example 1 continued** Suppose there are  $n$  lenders. As in the original example, each lender has vast, distinct data on group  $A = 1$ , and trains a classifier to predict loan repayment. No lender has sufficient data on group  $A = 0$ , so both outsource to a third-party, who provides each lender with an (identical) classifier. Suppose all classifiers have the same false-negative rate  $\beta$ . Thus, a firm that uses its own classifier on  $A = 1$  and the third-party classifier on  $A = 0$  is in practice using an EO classifier. Suppose also that, on  $A = 1$ , the firms' individual classifiers misclassify positive instances independently of other classifiers' predictions:

$$\Pr[c_1(X, A) = \dots = c_n(X, A) = 0|A = 1, Y = 1] = \beta^n.$$

Additionally, since all lenders use the same classifier on  $A = 0$ , the probability that a positive instance from  $A = 0$  is misclassified,  $\Pr[c_1(X, A) = \dots = c_n(X, A) = 0|A = 0, Y = 1]$ , is equal to  $\beta$ . Thus, even though each firm uses an EO classifier, the level of EOC is  $\beta - \beta^n$ . Observe that this is increasing in  $n$ .

The main insight from this section is that rather than improving the situation, increasing the number of (EO) classifiers can actually lead to a larger EOC.

## B Proof of Lemma 1

Recall the assumption that the fairness adjustment is derived via post-processing (Hardt et al., 2016): Given a learned classifier  $c$ , the EO classifier  $\tilde{c}$  is *derived* from  $c$ , namely, it depends only on  $A$  and the predictions  $c(X, A)$ . We also assume that  $\tilde{c}$  minimizes squared-loss relative to all such derived EO classifiers. Under these simplifying assumptions we prove the following lemma:

**Lemma 1** Fix a classifier with false-negative rates  $\beta^0$  on group  $A = 0$  and  $\beta^1$  on group  $A = 1$ . Then for each  $\beta \in \{\beta^0, \beta^1\}$  there exists a distribution  $\mathcal{D}$  and false positive-rates under which the optimal derived EO classifier has false-negative rate  $\beta$  on both groups.

**Proof:** For a classifier  $c$  and a group  $a \in A$ , let  $\lambda^a(c) = (\alpha^a, 1 - \beta^a)$ , where  $\alpha^a = \Pr[c_L(X, A, Y) = 1 | Y = 0, A = a]$  is the false-positive rate of the classifier in group  $a$ , and  $1 - \beta^a = \Pr[c_L(X, A, Y) = 1 | Y = 1, A = a]$  is the true-positive rate of the classifier in group  $a$ . Hardt et al. (2016) show that any derived classifier  $\tilde{c}$  satisfies  $\lambda^a(\tilde{c}) \in \text{convhull}\{(0, 0), \lambda^a(c), \lambda^a(1 - c), (1, 1)\}$ , where  $(1 - c)$  is the classifier  $c$  but with predictions flipped. They then show that the optimal derived classifier can be found using the following linear program:

$$\begin{aligned} \min_{\tilde{c}} \quad & \mathbb{E}[\ell(\tilde{c}, c)] \\ \text{s.t.} \quad & \lambda^a(\tilde{c}) \in \text{convhull}\{(0, 0), \lambda^a(c), \lambda^a(1 - c), (1, 1)\}, \quad \forall a \in A \\ & \lambda_2^0(\tilde{c}) = \lambda_2^1(\tilde{c}) \end{aligned}$$

The second constraint above states that the true-positive rates of the classifier are equal on both  $a \in A$ , and so the classifier  $\tilde{c}$  must be EO. The optimization problem is a squared-loss minimization.

Let us fix the distribution  $\mathcal{D}$  to be uniform over all of  $X \times Y \times A$ . Thus, for each group  $a$  and label  $y \in \{0, 1\}$ , we have that  $\Pr[A = a] = \Pr[Y = y] = 1/2$ . We also assume, without loss of generality, that  $\beta^0 > \beta^1$ .

The optimal derived classifier  $\tilde{c}$  is then one of the following:

1. Let  $\tilde{c} \equiv c$  except that, on some fraction of instances with  $c(x, 0) = 0$  fix  $\tilde{c}(x, 0) = 1$ , so that

$$\lambda^0(\tilde{c}) = (1 - \delta^0)\lambda^0(c) + \delta^0 \cdot (1, 1) = (\tilde{\alpha}^0, 1 - \beta^1)$$

for  $\delta^0 = (\beta^0 - \beta^1)/\beta^0$  and some  $\tilde{\alpha}^0$ . In this case,  $\tilde{\beta}^0 = \beta^1$ . The cost of this in terms of squared loss is the fraction of instances where  $c(x, 0) = 0$  that were flipped to  $\tilde{c}(x, 0) = 1$ , namely,  $\delta^0(1 - \alpha^0 + \beta^0)/4$ .

2. Let  $\tilde{c} \equiv c$  except that, on some fraction of instances with  $c(x, 1) = 1$  fix  $\tilde{c}(x, 1) = 0$ , so that

$$\lambda^1(\tilde{c}) = (1 - \delta^1)\lambda^1(c) + \delta^1 \cdot (0, 0) = (\tilde{\alpha}^1, 1 - \beta^0)$$

for  $\delta^1 = (\beta^0 - \beta^1)/(1 - \beta^1)$  and some  $\tilde{\alpha}^1$ . In this case,  $\tilde{\beta}^1 = \beta^0$ . The cost of this in terms of squared loss is the fraction of instances where  $c(x, 1) = 1$  that were flipped to  $\tilde{c}(x, 1) = 0$ , namely,  $\delta^1(\alpha^1 + 1 - \beta^0)/4$ .

To see that one of the above is an optimal derived classifier, note first that the optimal solution cannot be in the strict interior of the polytope. Furthermore, if  $(1 - \alpha^0 + \beta^0)/\beta^0 \neq (\alpha^1 + 1 - \beta^0)/(1 - \beta^1)$  then considering a mixture of (1) and (2)—on some fraction of instances with  $c(x, 0) = 0$  setting  $\tilde{c}(x, 0) = 1$  and on some fraction of instances with  $c(x, 1) = 1$  setting  $\tilde{c}(x, 1) = 0$ —is suboptimal. Finally, note that setting some fraction of instances with  $c(x, 0) = 1$  to  $\tilde{c}(x, 0) = 0$  only increases the EO, as does setting some fraction of instances with  $c(x, 1) = 0$  to  $\tilde{c}(x, 1) = 1$ , and so cannot be part of an optimal classifier.

Finally, an optimal derived EO classifier in which the false-negative rate is  $\beta^1$  (resp.,  $\beta^0$ ) on both groups is one where false-positive rates satisfy  $(1 - \alpha^0 + \beta^0)/\beta^0 < (\alpha^1 + 1 - \beta^0)/(1 - \beta^1)$  (resp.,  $(1 - \alpha^0 + \beta^0)/\beta^0 > (\alpha^1 + 1 - \beta^0)/(1 - \beta^1)$ ). Under equality, both classifiers are optimal. ■

## References

- HARDT, M., PRICE, E. and SREBRO, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, **29**.