On Fairness and Calibration

Recently, there has been growing concern about errors of machine learning algorithms in sensitive domains—including criminal justice, online advertising, and medical testing [Executive Office of the President 2016]—that may systematically discriminate against particular groups of people [Barocas and Selbst 2016, Chouldechova 2017, Berk et al. 2018].

A recent high-profile example of these concerns was raised by the news organization ProPublica, who studied a risk-assessment tool that is widely used in the criminal justice system. This tool assigns to each criminal defendant an estimated probability that they will commit a future crime. ProPublica found that the risk estimates assigned to defendants who did not commit future crimes were on average higher among African American defendants than Caucasian defendants [Angwin et al. 2022]. This is a form of false-positive error, and in this case it disproportionately affected African American defendants. To mitigate issues such as these, the machine learning community has proposed different frameworks that attempt to quantify fairness in classification [Barocas and Selbst 2016, Hardt et al. 2016b, Chouldechova 2017, Kleinberg et al. 2017, Woodworth et al. 2017, Zafar et al. 2017a, Berk et al. 2018]. A recent and particularly noteworthy framework is Equalized Odds [Hardt et al. 2016b] (also referred to as Disparate Mistreatment [Zafar et al. 2017a]),¹ which constrains classification algorithms in such a way that no error type (false-positive or false-negative) disproportionately affects any population subgroup. This notion of non-discrimination is feasible in many settings, and researchers have developed tractable algorithms for achieving it [Goh et al. 2016, Hardt et al. 2016b, Woodworth et al. 2017, Zafar et al. 2017a].

When risk tools are used in practice, a key goal is that they are calibrated: if we look at the set of people who receive a predicted probability of \( p \), we would like

---

¹. For the remainder of the chapter, we will use *Equalized Odds* to refer to this notion of non-discrimination.
a \( p \) fraction of the members of this set to be positive instances of the classification problem [Dawid 1982]. Moreover, if we are concerned about fairness between two groups \( G_1 \) and \( G_2 \) (e.g., African American defendants and white defendants), then we would like this calibration condition to hold simultaneously for the set of people within each of these groups as well [Flores et al. 2016].

Calibration is a crucial condition for risk tools in many settings. If a risk tool for evaluating defendants were not calibrated with respect to groups defined by race, for example, then a probability estimate of \( p \) could carry different meaning for African American and white defendants, and hence the tool would have the unintended and highly undesirable consequence of incentivizing judges to take race into account when interpreting its predictions.

Despite the importance of calibration as a property, our understanding of how it interacts with other fairness properties is limited. We know from recent work that, except in the most constrained cases, it is impossible to achieve calibration while also satisfying Equalized Odds [Chouldechova 2017, Kleinberg et al. 2017]. However, we do not know how best to achieve relaxations of these guarantees that are feasible in practice.

Our goal is to further investigate the relationship between calibration and error rates. We show that even if the Equalized Odds conditions are relaxed substantially—requiring only that weighted sums of the group error rates match—it is still problematic to also enforce calibration. We provide necessary and sufficient conditions under which this calibrated relaxation is feasible. When feasible, it has a unique optimal solution that can be achieved through post-processing of existing classifiers. Moreover, we provide a simple post-processing algorithm to find this solution: withholding predictive information for randomly chosen inputs to achieve parity and preserve calibration. However, this simple post-processing method is fundamentally unsatisfactory: although the post-processed predictions of our information-withholding algorithm are “fair” in expectation, most practitioners would object to the fact that a non-trivial portion of the individual predictions are withheld as a result of coin tosses—especially in sensitive settings such as health care or criminal justice. The optimality of this algorithm thus has troubling implications and shows that calibration and error-rate fairness are inherently at odds (even beyond the initial results by Chouldechova [2017] and Kleinberg et al. [2017]).

Finally, we evaluate these theoretical findings empirically, comparing calibrated notions of non-discrimination against the (uncalibrated) Equalized Odds framework on several datasets. These experiments further support our conclusion that calibration and error-rate constraints are in most cases mutually incompatible.
goals. In practical settings, it may be advisable to choose only one of these goals rather than attempting to achieve some relaxed notion of both.

### 2.1 Problem Setup

The setup of our framework most follows the Equal Odds framework [Hardt et al. 2016b, Zafar et al. 2017a]; however, we extend their framework for use with probabilistic classifiers. Let $P \subseteq \mathbb{R}^k \times \{0, 1\}$ be the input space of a binary classification task. In our criminal justice example, $(x, y) \sim P$ represents a person, with $x$ representing the individual’s history and $y$ representing whether or not the person will commit another crime. Additionally, we assume the presence of two groups $G_1, G_2 \subseteq P$, which represent disjoint population subsets, such as different races. We assume that the groups have different base rates $\mu_t$, or probabilities of belonging to the positive class: $\mu_1 = \text{Pr}_{(x,y) \sim G_1}[y = 1] \neq \text{Pr}_{(x,y) \sim G_2}[y = 1] = \mu_2$.

Finally, let $h_1, h_2 : \mathbb{R}^k \to \{0, 1\}$ be binary classifiers, where $h_1$ classifies samples from $G_1$ and $h_2$ classifies samples from $G_2$. Each classifier outputs the probability that a given sample $x$ belongs to the positive class. The notion of Equalized Odds non-discrimination is based on the false-positive and false-negative rates for each group, which we generalize here for use with probabilistic classifiers:

**Definition 2.1** The generalized false-positive rate of classifier $h_t$ for group $G_t$ is $c_{fp}(h_t) = \mathbb{E}_{(x,y) \sim G_t}[h_t(x) \mid y = 0]$. Similarly, the generalized false-negative rate of classifier $h_t$ is $c_{fn}(h_t) = \mathbb{E}_{(x,y) \sim G_t}[(1 - h_t(x)) \mid y = 1]$.

If the classifier were to output either 0 or 1, this represents the standard notions of false-positive and false-negative rates. We now define the Equalized Odds framework (generalized for probabilistic classifiers), which aims to ensure that errors of a given type are not biased against any group.

**Definition 2.2** **Probabilistic Equalized Odds**

Classifiers $h_1$ and $h_2$ exhibit Equalized Odds for groups $G_1$ and $G_2$ if $c_{fp}(h_1) = c_{fp}(h_2)$ and $c_{fn}(h_1) = c_{fn}(h_2)$.

**Calibration constraints.** As stated in the introduction, these two conditions do not necessarily prevent discrimination if the classifier predictions do not represent well-calibrated probabilities. Recall that calibration intuitively says that probabilities should carry semantic meaning: if there are 100 people in $G_1$ for whom $h_1(x) = 0.6$, then we expect 60 of them to belong to the positive class.

**Definition 2.3** A classifier $h_t$ is perfectly calibrated if $\forall p \in [0, 1], \text{Pr}_{(x,y) \sim G_t}[y = 1 \mid h_t(x) = p] = p$.  

2. In practice, $h_1$ and $h_2$ can be trained jointly (i.e., they are the same classifier).
It is commonly accepted amongst practitioners that both classifiers $h_1$ and $h_2$ should be calibrated with respect to groups $G_1$ and $G_2$ to prevent discrimination [Crowson et al. 2016, Dieterich et al. 2016, Flores et al. 2016, Berk et al. 2018]. Intuitively, this prevents the probability scores from carrying group-specific information. Unfortunately, Kleinberg et al. [2017] (as well as Chouldechova [2017], in a binary setting) prove that a classifier cannot achieve both calibration and Equalized Odds, even in an approximate sense, except in the most trivial of cases.

### 2.1.1 Geometric Characterization of Constraints

We now will characterize the calibration and error-rate constraints with simple geometric intuitions. Throughout the rest of this chapter, all of our results can be easily derived from this interpretation. We begin by defining the region of classifiers that are trivial, or those that output a constant value for all inputs (i.e., $h^c(x) = c$, where $0 \leq c \leq 1$ is a constant). We can visualize these classifiers on a graph with generalized false-positive rates on one axis and generalized false-negatives on the other. It follows from the definitions of generalized false-positive/false-negative rates and calibration that all trivial classifiers $h$ lie on the diagonal defined by $c_{FP}(h) + c_{FN}(h) = 1$ (Figure 2.1(a)). Therefore, all classifiers that are “better than random” must lie below this diagonal in false-positive/false-negative space (the gray triangle in the figure). Any classifier that lies above the diagonal performs “worse than random,” as we can find a point on the trivial classifier diagonal with lower false-positive and false-negative rates.

Now we will characterize the set of calibrated classifiers for groups $G_1$ and $G_2$, which we denote as $H_1^c$ and $H_2^c$. Kleinberg et al. [2017] show that the generalized false-positive and false-negative rates of a calibrated classifier are linearly related.

![Figure 2.1](image)

**Figure 2.1** Calibration, trivial classifiers, and equal-cost constraints—plotted in the false-positive (FP)/false-negative (FN) plane. $H_1^c$, $H_2^c$ are the set of calibrated classifiers for the two groups, and $h^{c_1}, h^{c_2}$ are trivial classifiers. (a) Possible calibrated classifiers $H_1^c, H_2^c$ (blue/red). (b) Satisfying calibrated and equal false-positive rates. (c) Satisfying calibrated and equal false-negative rates. (d) Satisfying calibration and a general constraint.
by the base rate of the group:

\[ c_{fn}(h) = \frac{(1 - \mu)}{\mu} c_{fp}(h). \] (2.1)

In other words, \( h_1 \) lies on a line with slope \( (1 - \mu_1)/\mu_1 \) and \( h_2 \) lies on a line with slope \( (1 - \mu_2)/\mu_2 \) (Figure 2.1(a)). The lower endpoint of each line is the perfect classifier, which assigns the correct prediction with complete certainty to every input. The upper endpoint is a trivial classifier as no calibrated classifier can perform “worse than random” (see Lemma B.2 in Section B.2). The only trivial classifier that satisfies the calibration condition for a group \( G_t \) is the one that outputs the base rate \( \mu_t \). We will refer to \( h^{\mu_1} \) and \( h^{\mu_2} \) as the trivial classifiers, calibrated for groups \( G_1 \) and \( G_2 \), respectively. It follows from the definitions that \( c_{fp}(h^{\mu_1}) = \mu_1 \) and \( c_{fn}(h^{\mu_2}) = 1 - \mu_1 \), and likewise for \( h^{\mu_2} \).

Finally, it is worth noting that for calibrated classifiers a lower false-positive rate necessarily corresponds to a lower false-negative rate and vice-versa. In other words, for a given base rate, a “better” calibrated classifier lies closer to the origin on the line of calibrated classifiers.

**Impossibility of equalized odds with calibration.** With this geometric intuition, we can provide a simplified proof of the main impossibility result from Kleinberg et al. [2017]:

**Theorem**

**Impossibility Result** [Kleinberg et al. 2017]

Let \( h_1 \) and \( h_2 \) be classifiers for groups \( G_1 \) and \( G_2 \) with \( \mu_1 \neq \mu_2 \). \( h_1 \) and \( h_2 \) satisfy the Equalized Odds and calibration conditions if and only if \( h_1 \) and \( h_2 \) are perfect predictors.

Intuitively, the three conditions define a set of classifiers that is overconstrained. Equalized Odds stipulates that the classifiers \( h_1 \) and \( h_2 \) must lie on the same coordinate in the false-positive/false-negative plane. As \( h_1 \) must lie on the blue line of calibrated classifiers for \( \mathcal{H}^+_1 \) and \( h_2 \) on the red line \( \mathcal{H}^+_2 \), they can only satisfy EO at the unique intersection point—the origin (and location of the perfect classifier). This implies that unless the two classifiers achieve perfect accuracy, we must relax the Equalized Odds conditions if we want to maintain calibration.

### 2.2 Relaxing Equalized Odds to Preserve Calibration

In this section, we show that a substantially simplified notion of Equalized Odds is compatible with calibration. We introduce a general relaxation that seeks to satisfy a single equal-cost constraint while maintaining calibration for each group \( G_t \).

---

3. Throughout this chapter we will treat the calibration constraint as holding exactly; however, our results generalize to approximate settings as well. See Appendix B for more details.
We begin with the observation that Equalized Odds sets constraints to equalize false-positives \( c_{fp}(h_i) \) and false-negatives \( c_{fn}(h_i) \). To capture and generalize this, we define a cost function \( g_t \) to be a linear function in \( c_{fp}(h_i) \) and \( c_{fn}(h_i) \) with arbitrary dependence on the group’s base rate \( \mu_t \). More formally, a cost function for group \( G_t \) is

\[
g_t(h_i) = a_t c_{fp}(h_i) + b_t c_{fn}(h_i),
\]

where \( a_t \) and \( b_t \) are non-negative constants that are specific to each group (and thus may depend on \( \mu_t \)); see Figure 2.1(d). We also make the assumption that for any \( \mu_t \), at least one of \( a_t \) and \( b_t \) is non-zero, meaning \( g_t(h_i) = 0 \) if and only if \( c_{fp}(h_i) = c_{fn}(h_i) = 0 \). This class of cost functions encompasses a variety of scenarios. As an example, imagine an application in which the equal false-positive condition is essential but not the false-negative condition. Such a scenario may arise in our recidivism-prediction example if we require that non-repeat offenders of any race be not disproportionately labeled as high risk. If we plot the set of calibrated classifiers \( H_t^1 \) and \( H_t^2 \) on the false-positive/false-negative plane, we can see that ensuring the false-positive condition requires finding classifiers \( h_1 \in H_t^1 \) and \( h_2 \in H_t^2 \) that fall on the same vertical line (Figure 2.1(b)). Conversely, if we instead choose to satisfy only the false-negative condition, we would find classifiers \( h_1 \) and \( h_2 \) that fall on the same horizontal line (Figure 2.1(c)). Finally, if both false-positive and false-negative errors incur a negative cost on the individual, we may choose to equalize a weighted combination of the error rates [Berk 2016, Chouldechova 2017, Berk et al. 2018], which can be graphically described by the classifiers lying on a convex and negatively sloped level set (Figure 2.1(d)). With these definitions, we can formally define our relaxation:

**Definition 2.4** Relaxed Equalized Odds with Calibration

Given a cost function \( g_t \) of the form in (2.2), classifiers \( h_1 \) and \( h_2 \) achieve Relaxed Equalized Odds with Calibration for groups \( G_1 \) and \( G_2 \) if both classifiers are calibrated and satisfy the constraint \( g_1(h_1) = g_2(h_2) \).

It is worth noting that, for calibrated classifiers, an increase in cost strictly corresponds to an increase in both the false-negative and false-positive rate. This can be interpreted graphically, as the level-order cost curves lie further away from the origin as cost increases (Figure 2.2(a)). In other words, the cost function can always be used as a proxy for either error rate.\(^5\)

---

4. By calibration, we cannot have one of \( c_{fp}(h_i) = 0 \) or \( c_{fn}(h_i) = 0 \) without the other, see Figure 2.1(a).

5. This holds even for approximately calibrated classifiers—see Section B.3.
Feasibility. It is easy to see that Definition 2.4 is always satisfiable—in Figure 2.1(b), (c), and (d) we see that there are many such solutions that would lie on a given level-order cost curve while maintaining calibration, including the case in which both classifiers are perfect. In practice, however, not all classifiers are achievable. For the rest of the chapter, we will assume that we have access to “optimal” (but possibly discriminatory) calibrated classifiers $h_1$ and $h_2$ such that, due to whatever limitations there are on the predictability of the task, we are unable to find other classifiers that have lower cost with respect to $g_t$. We allow $h_1$ and $h_2$ to be learned in any way as long as they are calibrated. Without loss of generality, for the remainder of the chapter, we will assume that $g_1(h_1) \geq g_2(h_2)$.

Since by assumption we have no way to find a classifier for $G_1$ with lower cost than $h_1$, our goal is therefore to find a classifier $\hat{h}_2$ with cost equal to $h_1$. This pair of classifiers would represent the lowest cost (and therefore optimal) set of classifiers that satisfies calibration and the equal cost constraint. For a given base rate $\mu_t$ and value of the cost function $g_t$, a calibrated classifier’s position in the generalized false-positive/false-negative plane is uniquely determined (Figure 2.2(a)). This is because each level-order curve of the cost function $g_t$ has negative slope in this plane, and each level order curve only intersects a group’s calibrated classifier line once. In other words, there is a unique solution in the false-positive/false-negative plane for classifier $\hat{h}_2$ (Figure 2.2(b)).

Consider the range of values that $g_t$ can take. As noted above, $g_t(h_1) \geq 0$, with equality if and only if $h_t$ is the perfect classifier. On the other hand, the trivial classifier (again, which outputs the constant $\mu_t$ for all inputs) is the calibrated classifier that achieves maximum cost for any $g_t$ (see Lemma B.2 in Section B.2). As a result,

![Figure 2.2](image)

**Figure 2.2** Calibration-Preserving Parity through interpolation. (a) Level-order curves of cost. Low cost implies low error rates. (b) Usually, there is a calibrated classifier $\hat{h}_2$ with the same cost of $h_1$. (c) Calibration and equal cost are incompatible if $h_1$ has high error. (d) Possible calibrated classifiers for $G_2$ (bold red) by mixing $h_2$ and $h^{h_2}$. FP: false positive; FN: false negative.
the cost of a classifier for group $G_\ell$ is between 0 and $g_\ell(h^{\mu_\ell})$. This naturally leads to a characterization of feasibility: Definition 2.4 can be achieved if and only if $h_1$ incurs less cost than group $G_2$’s trivial classifier $h^{\mu_2}$; that is, if $g_1(h_1) \leq g_2(h^{\mu_2})$. This can be seen graphically in Figure 2.2(c), in which the level-order curve for $g_1(h_1)$ does not intersect the set of calibrated classifiers for $G_2$. Since, by assumption, we cannot find a calibrated classifier for $G_1$ with strictly smaller cost than $h_1$, there is no feasible solution. On the other hand, if $h_1$ incurs less cost than $h^{\mu_2}$, then we will show feasibility by construction with a simple algorithm.

**An algorithm.** While it may be possible to encode the constraints of Definition 2.4 into the training procedure of $h_1$ and $h_2$, it is not immediately obvious how to do so. Even naturally probabilistic algorithms, such as logistic regression, can become uncalibrated in the presence of optimization constraints (as is the case in Zafar et al. [2017a]). It is not straightforward to encode the calibration constraint if the probabilities are assumed to be continuous, and post-processing calibration methods [Platt 1999, Zadrozny and Elkan 2001] would break equal-cost constraints by modifying classifier scores. Therefore, we look to achieve the calibrated Equalized Odds relaxation by post-processing existing calibrated classifiers.

Again, given $h_1$ and $h_2$ with $g_1(h_1) \geq g_2(h_2)$, we want to arrive at a calibrated classifier $\tilde{h}_2$ for group $G_2$ such that $g_1(h_1) = g_2(\tilde{h}_2)$. Recall that, under our assumptions, this would be the best possible solution with respect to classifier cost. We show that this cost constraint can be achieved by withholding predictive information for a randomly chosen subset of group $G_2$. In other words, rather than always returning $h_2(x)$ for all samples, we will occasionally return the group’s mean probability (i.e., the output of the trivial classifier $h^{\mu_2}$). In Lemma B.3 in Section B.2, we show that if

$$\tilde{h}_2(x) = \begin{cases} h^{\mu_2}(x) = \mu_2 & \text{with probability } \alpha \\ h_2(x) & \text{with probability } 1 - \alpha \end{cases} \quad (2.3)$$

then the cost of $\tilde{h}_2$ is a linear interpolation between the costs of $h_2$ and $h^{\mu_2}$ (Figure 2.2(d)). More formally, we have that $g_2(\tilde{h}_2) = (1 - \alpha)g_2(h_2) + \alpha g_2(h^{\mu_2})$, and thus setting $\alpha = \frac{g_1(h_1) - g_2(h_2)}{g_2(h^{\mu_2}) - g_2(h_2)}$ ensures that $g_2(\tilde{h}_2) = g_1(h_1)$ as desired (Figure 2.2(b)). Moreover, this randomization preserves calibration (see Section B.4).

Algorithm 2.1 summarizes this method.

**Implications.** In a certain sense, Algorithm 2.1 is an “optimal” method because it arrives at the unique false-negative/false-positive solution for $\tilde{h}_2$, where $\tilde{h}_2$ is calibrated and has cost equal to $h_1$. Therefore (by our assumptions), we can find no better classifiers that satisfy Definition 2.4. This simple result has strong consequences, as the tradeoffs to satisfy both calibration and the equal-cost constraint
Algorithm 2.1: Achieving Calibration and an Equal-Cost Constraint via Information Withholding

**Input:** classifiers $h_1$ and $h_2$ s.t. $g_2(h_2) \leq g_1(h_1) \leq g_2(h_2)$, holdout set $P_{\text{valid}}$.  
- Determine base rate $\mu_2$ of $G_2$ (using $P_{\text{valid}}$) to produce trivial classifier $h^{\mu_2}$.  
- Construct $\tilde{h}_2$ using with $\alpha = \frac{g_2(h_1) - g_2(h_2)}{g_2(h^{\mu_2}) - g_2(h_2)}$, where $\alpha$ is the interpolation parameter.  

**return** $h_1, \tilde{h}_2$ — which are calibrated and satisfy $g_1(h_1) = g_2(\tilde{h}_2)$.

are often unsatisfactory—both intuitively and experimentally (as we will show in Section 2.3).

We find two primary objections to this solution. First, it equalizes costs simply by making a classifier strictly worse for one of the groups. Second, it achieves this cost increase by withholding information on a randomly chosen population subset, making the outcome inequitable within the group (as measured by a standard measure of inequality like the Gini coefficient). Due to the optimality of the algorithm, the former of these issues is unavoidable in any solution that satisfies Definition 2.4. The latter, however, is slightly more subtle and brings up the question of individual fairness (what guarantees we would like an algorithm to make with respect to each individual) and how it interacts with group fairness (population-level guarantees). While this certainly is an important issue for future work, in this particular setting, even if one could find another algorithm that distributes the burden of additional cost more equitably, any algorithm will make at least as many false-positive/false-negative errors as Algorithm 2.1, and these misclassifications will always be tragic to the individuals whom they affect. The performance loss across the entire group is often significant enough to make this combination of constraints somewhat worrying to use in practice, regardless of the algorithm.

**Impossibility of satisfying multiple equal-cost constraints.** It is natural to argue there might be multiple cost functions that we would like to equalize across groups. However, satisfying more than one distinct equal-cost constraint (i.e., different curves in the false-positive/false-negative plane) is infeasible.

**Theorem 2.1** Generalized impossibility result

Let $h_1$ and $h_2$ be calibrated classifiers for $G_1$ and $G_2$ with equal cost with respect to $g_t$. If $\mu_1 \neq \mu_2$, and if $h_1$ and $h_2$ also have equal cost with respect to a different cost function $g'_t$, then $h_1$ and $h_2$ must be perfect classifiers.

(Proof in Section B.5). Note that this is a generalization of the impossibility result of Kleinberg et al. [2017]. Furthermore, we show in Theorem B.3 (in
Section B.5) that this holds in an approximate sense: if calibration and multiple distinct equal-cost constraints are approximately achieved by some classifier, then that classifier must have approximately zero generalized false-positive and false-negative rates.

### 2.3 Experiments

In light of these findings, our goal is to understand the impact of imposing calibration and an equal-cost constraint on real-world datasets. We will empirically show that, in many cases, this will result in performance degradation while simultaneously increasing other notions of disparity. We perform experiments on three datasets: an income-prediction, a health-prediction, and a criminal recidivism dataset. For each task, we choose a cost function within our framework that is appropriate for the given scenario. We begin with two calibrated classifiers \( h_1 \) and \( h_2 \) for groups \( G_1 \) and \( G_2 \). We assume that these classifiers cannot be significantly improved without more training data or features. We then derive \( \hat{h}_2 \) to equalize the costs while maintaining calibration. The original classifiers are trained on a portion of the data, and then the new classifiers are derived using a separate holdout set. To compare against the (uncalibrated) Equalized Odds framework, we derive false-positive/false-negative matching classifiers using the post-processing method of Hardt et al. [2016b] (EO-Derived). On the criminal recidivism dataset, we additionally learn classifiers that directly encode the Equalized Odds constraints, using the methods of Zafar et al. [2017a] (EO-Trained). (See Section B.6 for detailed training and post-processing procedures.) We visualize model error rates on the generalized false-positive and false-negative plane. Additionally, we plot the calibrated classifier lines for \( G_1 \) and \( G_2 \) to visualize model calibration.

**Income prediction.** The Adult Dataset from UCI Machine Learning Repository [Lichman 2013] contains 14 demographic and occupational features for various people, with the goal of predicting whether a person’s income is above $50,000. In this scenario, we seek to achieve predictions with equalized cost across genders (\( G_1 \) represents women and \( G_2 \) represents men). We model a scenario where the primary concern is ensuring equal generalized false-negative rates across genders, which would, for example, help job recruiters prevent gender discrimination in the form of underestimated salaries. Thus, we choose our cost constraint to require equal generalized false-negative rates across groups. In Figure 2.3(a), we see that the original classifiers \( h_1 \) and \( h_2 \) approximately lie on the line of calibrated classifiers. In the left plot (EO-Derived), we see that it is possible to (approximately) match both error rates of the classifiers at the cost of \( h_1^{\text{EO}} \) deviating from the set of calibrated classifiers. In the right plot, we see that it is feasible to equalize the
2.3 Experiments

![Diagram showing generalized false-positive (FP) and false-negative (FN) rates for two groups under Equalized Odds and the calibrated relaxation. Diamonds represent post-processed classifiers. Points on the Equalized Odds (trained) graph represent classifiers achieved by modifying constraint hyperparameters. (a) Income Prediction. (b) Health Prediction. (c) Recidivism Prediction.]

Figure 2.3 Generalized false-positive (FP) and false-negative (FN) rates for two groups under Equalized Odds and the calibrated relaxation. Diamonds represent post-processed classifiers. Points on the Equalized Odds (trained) graph represent classifiers achieved by modifying constraint hyperparameters. (a) Income Prediction. (b) Health Prediction. (c) Recidivism Prediction.

generalized false-negative rates while maintaining calibration. \(h_1\) and \(h_2\) lie on the same level-order curve of \(g_t\) (represented by the dashed-gray line) and simultaneously remain on the “line” of calibrated classifiers. It is worth noting that achieving either notion of non-discrimination requires some cost to at least one of the groups. However, maintaining calibration further increases the difference in false-positive rates between groups. In some sense, the calibrated framework trades off one notion of disparity for another while simultaneously increasing the overall error rates.

**Health prediction.** The Heart Dataset from the UCI Machine Learning Repository contains 14 processed features from 906 adults in four geographical locations. The goal of this dataset is to accurately predict whether or not an individual has a heart condition. In this scenario, we would like to reduce disparity between middle-aged adults \((G_1)\) and seniors \((G_2)\). In this scenario, we consider false-positive and false-negative to both be undesirable. A false prediction of a heart condition could result in unnecessary medical attention, while false negatives incur cost from delayed
treatment. We therefore utilize the following cost function \( g(h) = r_f p h(x)(1 - y) + r_f n(1 - h(x)) y \), which essentially assigns a weight to both false-negative and false-positive predictions. In our experiments, we set \( r_{fp} = 1 \) and \( r_{fn} = 3 \). In the right plot of Figure 2.3(b), we can see that the level-order curves of the cost function form a curved line in the generalized false-positive/false-negative plane. Because our original classifiers lie approximately on the same level-order curve, little change is required to equalize the costs of \( h_1 \) and \( \tilde{h}_2 \) while maintaining calibration. This is the only experiment in which the calibrated framework incurs little additional cost, and therefore could be considered a viable option. However, it is worth noting that, in this example, the equal-cost constraint does not explicitly match either of the error types, and therefore the two groups will in expectation experience different types of errors. In the left plot of Figure 2.3(b) (EO-Derived), we see that it is alternatively feasible to explicitly match both the false-positive and false-negative rates while sacrificing calibration.

**Criminal recidivism prediction.** Finally, we examine the frameworks in the context of our motivating example: criminal recidivism. As mentioned in the introduction, African Americans \( (G_1) \) receive a disproportionate number of false-positive predictions as compared with Caucasians \( (G_2) \) when automated risk tools are used in practice. Therefore, we aim to equalize the generalized false-positive rate. In this experiment, we modify the predictions made by the COMPAS tool [Dieterich et al. 2016], a risk-assessment tool used in practice by the American legal system. Additionally, we also see if it is possible to improve the classifiers with training-time Equalized Odds constraints using the methods of Zafar et al. [2017a] (EO-Trained). In Figure 2.3(c), we first observe that the original classifiers \( h_1 \) and \( h_2 \) have large generalized false-positive and false-negative rates. Both methods of achieving Equalized Odds—training constraints (left plot) and post-processing (middle plot)—match the error rates while sacrificing calibration. However, we observe that, assuming \( h_1 \) and \( h_2 \) cannot be improved, it is infeasible to achieve the calibrated relaxation (Figure 2.3(c) right). This is an example where matching the false-positive rate of \( h_1 \) would require a classifier worse than the trivial classifier \( h^{\mu_2} \). This example therefore represents an instance in which calibration is completely incompatible with any error-rate constraints. If the primary concern of criminal justice practitioners is calibration [Dieterich et al. 2016, Flores et al. 2016], then there will inherently be discrimination in the form of false-positive and false-negative rates. However, if the Equalized Odds framework is adopted, the miscalibrated risk scores inherently cause discrimination to one group, as argued in the introduction. Therefore, the most meaningful change in such a setting would be an improvement to \( h_2 \) (the classifier for African Americans) either through the collection of more data
or the use of more salient features. A reduction in overall error to the group with higher cost will naturally lead to less error-rate disparity.

2.4 Discussion and Conclusion

We have observed cases in which calibration and relaxed Equalized Odds are compatible and cases where they are not. When it is feasible, the penalty of equalizing cost is amplified if the base rates between groups differ significantly. This is expected since base rate differences are what give rise to cost disparity in the calibrated setting. Seeking equality with respect to a single error rate (e.g., false negatives, as in the income prediction experiment) will necessarily increase disparity with respect to the other error. This may be tolerable (in the income prediction case, some employees will end up overpaid) but could also be highly problematic (e.g., in criminal justice settings). Finally, we have observed that the calibrated relaxation is infeasible when the best (discriminatory) classifiers are not far from the trivial classifiers (leaving little room for interpolation). In such settings, we see that calibration is completely incompatible with an equalized error constraint.

In summary, we conclude that maintaining cost parity and calibration is desirable yet often difficult in practice. Although we provide an algorithm to effectively find the unique feasible solution to both constraints, it is inherently based on randomly exchanging the predictions of the better classifier with the trivial base rate. Even if fairness is reached in expectation, for an individual case, it may be hard to accept that occasionally consequential decisions are made by randomly withholding predictive information, irrespective of a particular person's feature representation. In this chapter, we argue that, as long as calibration is required, no lower-error solution can be achieved.