The Implicit Fairness Criterion of Unconstrained Learning

Lydia T. Liu [*1] **Max Simchowitz** [*1] **Moritz Hardt** [1]

Abstract

We clarify what fairness guarantees we can and cannot expect to follow from unconstrained machine learning. Specifically, we show that in many settings, unconstrained learning on its own implies *group calibration*, that is, the outcome variable is conditionally independent of group membership given the score. A lower bound confirms the optimality of our upper bound. Moreover, we prove that as the excess risk of the learned score decreases, the more strongly it violates *separation* and *independence*, two other standard fairness criteria. Our results challenge the view that group calibration necessitates an active intervention, suggesting that often we ought to think of it as a byproduct of unconstrained machine learning.

1. Introduction

Although many fairness-promoting interventions have been proposed in the machine learning literature, unconstrained learning remains the dominant paradigm among practitioners for learning risk scores from data. Given a prespecified class of models, unconstrained learning seeks to find a predictor which minimizes the average prediction loss over a labeled dataset, or some surrogate thereof, without explicitly correcting for disparity with respect to sensitive attributes, such as race or gender. Many criticize the practice of unconstrained machine learning for propagating harmful biases (Crawford, 2013; Barocas and Selbst, 2016; Crawford, 2017). Others see merit in unconstrained learning for reducing bias in consequential decisions (Kleinberg et al., 2016; Corbett-Davies et al., 2017a;b).

In this work, we show that defaulting to unconstrained learning does not neglect fairness considerations entirely. Instead, it prioritizes one notion of "fairness" over others: unconstrained learning achieves *calibration* with respect to one or more sensitive attributes, as well as a related criterion called *sufficiency* (e.g., Barocas et al., 2018), at the cost of violating other widely used fairness criteria, *separation* and *independence* (see Section 1.2 for references therein).

A risk score is *calibrated* for a group if the risk score obviates the need to solicit group membership for the purpose of predicting an outcome variable of interest. The concept of calibration has a venerable history in statistics and machine learning (Cox, 1958; Murphy and Winkler, 1977; Dawid, 1982; DeGroot and Fienberg, 1983; Platt, 1999; Zadrozny and Elkan, 2001; Niculescu-Mizil and Caruana, 2005). The appearance of calibration as a widely adopted and discussed "fairness criterion" largely resulted from a recent debate around fairness in recidivism prediction and pre-trial detention. After journalists at ProPublica pointed out that a popular recidivism risk score known as COMPAS had a disparity in false positive rates between white defendants and black defendants (Angwin et al., 2016), the organization that produced these scores countered that this disparity was a consequence of the fact that their scores were calibrated by race (Dieterich et al., 2016): that is, for both black and white individuals, the average probability of recidivism among of given score was nearly identical.

Formal trade-offs dating back the 1970s confirm the observed tension between calibration and other classification criteria, including the aforementioned criterion of *separation*, which is related to the disparity in false positive rates (Darlington, 1971; Chouldechova, 2017; Kleinberg et al., 2017; Barocas et al., 2018). In the context of COMPAS, for example, separation means that black and white individuals with similar probabilies of recidivism would receive similar scores.

Implicit in this debate is the view that calibration is a constraint that needs to be actively enforced as a means of promoting fairness. Consequently, recent literature has proposed new learning algorithms which ensure approximate calibration in different settings (Hebert-Johnson et al., 2018; Kearns et al., 2017).

The goal of this work is to understand when approximate calibration can in fact be achieved by unconstrained machine learning alone. We define several relaxations of the exact calibration criterion, and show that approximate group calibration is often a routine consequence of unconstrained

[*]Equal contribution [1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, USA. Correspondence to: Lydia T. Liu <lydiatliu@berkeley.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

learning. Such guarantees apply even when the sensitive attributes in question are not available to the learning algorithm. On the other hand, we demonstrate that under similar conditions, unconstrained learning strongly violates the *separation* and *independence* criteria. We also prove novel lower bounds which demonstrate that in the worst case, no other algorithm can produce score functions that are substantially better-calibrated than unconstrained learning. Finally, we verify our theoretical findings with experiments on two well-known datasets, demonstrating the effectiveness of unconstrained learning in achieving approximate calibration with respect to multiple group attributes simultaneously.

1.1. Summary of Results

We begin with a simplified presentation of our results. As is common in supervised learning, consider a pair of random variables (X, Y) where X models available features, and Y is a binary target variable that we try to predict from X. We choose a discrete random variable A in the same probability space to model group membership. For example, A could represent gender, or race. In general, X may include A, or features that are proxies for A. Our results *do not* require that X perfectly encodes the attribute A, and hence also apply to the setting where the sensitive attribute is unknown.

A score function f maps the random variable X to a real number. We say that the score function f is *sufficient* with respect to attribute A if we have $\mathbb{E}[Y \mid f(X)] = \mathbb{E}[Y \mid f(X), A]$ almost surely.[1] In words, conditioning on A provides no additional information about Y beyond what was revealed by $f(X)$. This definition leads to a natural notion of the *sufficiency gap*:

$$\mathbf{suf}_f(A) = \mathbb{E}[|\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y \mid f(X), A]|]\,, \quad (1)$$

which measures the expected deviation from satisfying sufficiency over a random draw of (X, A).

We say that the score function f is *calibrated* with respect to group A if we have $\mathbb{E}[Y \mid f(X), A] = f(X)$. Note that calibration implies sufficiency. We define the *calibration gap* (see also Pleiss et al., 2017) as

$$\mathbf{cal}_f(A) = \mathbb{E}\left[|f(X) - \mathbb{E}[Y \mid f(X), A]|\right]. \quad (2)$$

Denote by $\mathcal{L}(f) = \mathbb{E}[\ell(f, Y)]$ the *population risk* (*risk*, for short) of the score function f. Think of the loss function ℓ as either the square loss or the logistic loss, although our results apply more generally. Our first result relates the sufficiency and calibration gaps of a score to its risk.

[1]This notion has also been referred to as "calibration" in previous work (e.g., Chouldechova, 2017). In this work we refer to it as "sufficiency", hence distinguishing it from $\mathbb{E}[Y \mid f(X), A] = f(X)$, which has also been called "calibration" in previous work (e.g., Pleiss et al., 2017). These two notions are not identical, but closely related; we present analagous theoretical results for both.

Theorem 1.1 (Informal)**.** *For a broad class of loss functions that includes the square loss and logistic loss, we have*

$$\max\{\mathbf{suf}_f(A), \mathbf{cal}_f(A)\} \leq O\left(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}\right).$$

Here, $\mathcal{L}^*$ is the calibrated Bayes risk*, i.e., the risk of the score function $f^B(x, a) = \mathbb{E}[Y \mid X = x, A = a]$.*

The theorem shows that if we manage to find a score function with small excess risk over the calibrated Bayes risk, then the score function will also be reasonably sufficient and well-calibrated with respect to the group attribute A. We also provide analogous results for the calibration error restricted to a particular group $A = a$.

In particular, the above theorem suggests that computing the unconstrained *empirical risk minimizer* (Vapnik, 1992), or *ERM*, is a natural strategy for achieving group calibration and sufficiency. For a given loss $\ell : [0, 1] \times \{0, 1\} \to \mathbb{R}$, finite set of examples $S^n := \{(X_i, Y_i)\}_{i\in[n]}$, and class of possible scores $\mathcal{F}$, the ERM is the score function

$$\widehat{f}_n \in \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell(f(X_i), Y_i)\,. \quad (3)$$

It is well known that, under very general conditions, $\mathcal{L}(\widehat{f}_n) \overset{\text{prob}}{\to} \min_{f\in\mathcal{F}} \mathcal{L}(f)$; that is, the risk of $\widehat{f}_n$ converges in probability to the least expected loss of any score function $f \in \mathcal{F}$.

In general, the ERM may not achieve small excess risk, $\mathcal{L}(f) - \mathcal{L}^*$. Indeed, we have defined the calibrated Bayes score f^B as one that has access to both X and A. In cases where the available features X do not encode A, but A is relevant to the prediction task, the excess risk may be large. In other cases, the excess risk may be large simply because the function class over which we can feasibly optimize provides only poor approximations to the calibrated Bayes score. In example 2.1, we provide scenarios when the excess risk is indeed small.

The constant in front of the square root in our theorem depends on properties of the loss function, and is typically small, e.g., bounded by 4 for both the squared loss and the logistic loss. The more significant question is if the square root is necessary. We answer this question in the affirmative.

Theorem 1.2 (Informal)**.** *There is a triple of random variables (X, A, Y) such that the empirical risk minimizer $\widehat{f}_n$ trained on n samples drawn i.i.d. from (X, Y) satisfies $\min\{\mathbf{cal}_{\widehat{f}_n}(A), \mathbf{suf}_{\widehat{f}_n}(A)\} \geq \Omega(1/\sqrt{n})$ and $\mathcal{L}(\widehat{f}_n) - \mathcal{L}* \leq O(1/n)$ with probability $\Omega(1)$.*

In other words, our upper bound sharply characterizes the worst-case relationship between excess risk, sufficiency and calibration. Moreover, our lower bound applies not only to

the empirical risk minimizer $\widehat{f}_n$, but to any score learned from data which is a linear function of the features X. Although group calibration and sufficiency is a natural consequence of unconstrained learning, it is in general untrue that they imply a good predictor. For example, predicting the group average, $f = \mathbb{E}[Y \mid A]$ is a pathological score function that nevertheless satisfies calibration and sufficiency.

Although unconstrained learning leads to well-calibrated scores, it violates other notions of group fairness. We show that the ERM typically violates independence—the criterion that scores are independent of group attribute A—as long as the base rate $\Pr[Y = 1]$ differs by group. Moreover, we show that the ERM violates *separation*, which asks for scores $f(X)$ to be conditionally independent of the attribute A given the target Y (see Barocas et al., 2018, Chapter 2). In this work, we define the *separation gap*:

$$\mathbf{sep}_f(A) := \mathbb{E}_{Y,A}[|\mathbb{E}[f(X) \mid Y, A] - \mathbb{E}[f(X) \mid Y]|],$$

and show that any score with small excess risk must in general have a large separation gap. Similarly, we show that unconstrained learning violates $\mathbf{ind}_f(A) := \mathbb{E}_A[|\mathbb{E}[f(X) \mid A] - \mathbb{E}[f(X)]|]$, a quantitative version of the *independence* criterion (see Barocas et al., 2018, Chapter 2).

Theorem 1.3 (Informal)**.** *For a broad class of loss functions that includes the square loss and logistic loss, we have*

$$\mathbf{sep}_f(A) \geq C_{f_B} \cdot Q_A - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*}),$$

where C_{f_B} and Q_A are problem-specific constants independent of f. C_{f_B} represents the inherent noise level of the prediction task, and Q_A is the variation in group base rates. Moreover, $\mathbf{ind}_f(A) \geq Q_A - O(\sqrt{\mathcal{L}(f) - \mathcal{L}^*})$ for the same constant Q_A.

The lower bound for $\mathbf{sep}_f$ is explained in Section 2.2; the lower bound for $\mathbf{ind}_f$ is deferred to Appendix F.

Experimental evaluation. We explore the extent to which the result of empirical risk minimization satisfies sufficiency, calibration and separation, via comprehensive experiments on the UCI Adult dataset (Dua and Karra Taniskidou, 2017) and pretrial defendants dataset from Broward County, Florida (Angwin et al., 2016; Dressel and Farid, 2018). For various choices of group attributes, including those defined using arbitrary combinations of features, we observe that the empirical risk minimizing score is fairly close to being calibrated and sufficient. Notably, this holds even when the score is not a function of the group attribute in question.

1.2. Related Work

Calibration was first introduced as a fairness criterion by the education testing literature in the 1960s. It was formalized by the *Cleary criterion* (Cleary; 1968), which compares the slope of regression lines between the test score and the outcome in different groups. More recently, machine learning and data mining communities have rediscovered calibration, and examined the inherent tradeoffs between calibration and other fairness constraints. Chouldechova (2017) and Kleinberg et al. (2017) independently demonstrate that exact group calibration is incompatible with *separation* (equal true positive and false positive rates), except under highly restrictive situations such as perfect prediction or equal group base rates; further generalizations have been established by Pleiss et al. (2017).

There are multiple post-processing procedures which achieve calibration, (see e.g. Niculescu-Mizil and Caruana, 2005, and references therein). Notably, Platt scaling (Platt, 1999) learns calibrated probabilities for a given score function by logistic regression. Recently, Hebert-Johnson et al. (2018) proposed a polynomial time agnostic learning algorithm that achieves both low prediction error, and *multi-calibration*, or simultaneous calibration with respect to all, possibly overlapping, groups that can be described by a concept class of a given complexity. Complementary to this finding, our work shows that low prediction error often implies calibration with no additional computational cost, under very general conditions. Unlike Hebert-Johnson et al. (2018), we do not aim to guarantee calibration with respect to arbitrarily complex group structure; instead we study when usual empirical risk minimization already achieves calibration with respect to a given group attribute A.

A variety of other fairness criteria have been proposed to address concerns of fairness with respect to a sensitive attribute. These are typically group parity constraints on the score function, including, among others, *demographic parity* (also known as *independence* and *statistical parity*), *equalized odds* (also known as *error-rate balance* and *separation*), as well as *calibration* and *sufficiency* (see e.g. Feldman et al., 2015; Hardt et al., 2016; Chouldechova, 2017; Kleinberg et al., 2017; Pleiss et al., 2017; Barocas et al., 2018). Beyond parity constraints, recent works have also studied dynamic aspects of group fairness, such as the impact of model predictions on future welfare (Liu et al., 2018) and user demographics (Hashimoto et al., 2018). There are tensions between notions of group fairness and those of individual fairness (Dwork et al., 2012; Speicher et al., 2018). For a more complete treatment of algorithmic fairness literature, the reader is referred to Barocas et al. (2018); Chouldechova and Roth (2018); Corbett-Davies and Goel (2018).

2. Formal Setup and Results

We consider the problem of finding a *score function* $\widehat{f}$ which encodes the probability of a binary outcome $Y \in \{0, 1\}$, given access to features $X \in \mathcal{X}$. We consider functions

$f : \mathcal{X} \to [0,1]$ which lie in a prespecified function class $\mathcal{F}$. We assume that individuals' features and outcomes (X, Y) are random variables whose law is governed by a probability measure $\mathcal{D}$ over a space Ω, and will view functions f as maps $\Omega \to [0,1]$ via $f = f(X)$. We use $\Pr_{\mathcal{D}}[\cdot], \Pr[\cdot]$ to denote the probability of events under $\mathcal{D}$, and $\mathbb{E}_{\mathcal{D}}[\cdot], \mathbb{E}[\cdot]$ to denote expectation taken with respect to $\mathcal{D}$.

We also consider a $\mathcal{D}$-measurable protected attribute $A \in \mathcal{A}$, with respect to which we would like to ensure *sufficiency* or *calibration*, as defined in Section 1.1 above. While assume that $f = f(X)$ for all $f \in \mathcal{F}$, we compare the performance of f to the benchmark that we call the *calibrated Bayes score*[2]

$$f^B(x,a) := \mathbb{E}\left[Y \mid X = x, A = a\right], \tag{4}$$

which is a function of both the feature x and the attribute a. As a consequence, $f^B \notin \mathcal{F}$, except possibly whenever Y is conditionally independent of A given X. Nevertheless, f^B is well defined as a map $\Omega \to [0,1]$ and it always satisfies sufficiency and calibration:

Proposition 2.1. *f^B is sufficient and calibrated, that is $\mathbb{E}[Y \mid f^B(X)] = \mathbb{E}[Y \mid f^B(X), A]$ and $f^B = \mathbb{E}[Y \mid f(X), A]$, almost surely. Moreover, if $\Phi : \mathcal{X} \to \mathcal{X}'$ is any map, then the classifier $f_\Phi(X) := \mathbb{E}[Y \mid \Phi(X), A]$ is sufficient and calibrated.*

Proposition 2.1 is a direct consequence of the tower property (proof in Appendix A.1). In general, there are many challenges to learning perfectly calibrated scores. As mentioned above, f^B depends on information about A which is not necessarily accessible to scores $f \in \mathcal{F}$. Moreover, even in the setting where $A = A(X)$, it may still be the case that $\mathcal{F}$ is a restricted class of scores, and $f^B \notin \mathcal{F}$. Lastly, if $\widehat{f}$ is estimated from data, it may require infinitely many samples to achieve perfect calibration. To this end, we introduce the following approximate notion of sufficiency and calibration:

Definition 1. *Given a $\mathcal{D}$-measurable attribute $A \in \mathcal{A}$ and value $a \in \mathcal{A}$, we define the sufficiency gap of f with respect to A for group a, denoted $\mathbf{suf}_f(a; A)$, as*

$$\mathbb{E}_{\mathcal{D}}\left[\left|\mathbb{E}[Y \mid f(X)] - \mathbb{E}[Y \mid f(X), A]\right| \mid A = a\right] . \tag{5}$$

and the calibration gap for group a as

$$\mathbf{cal}_f(a; A) = \mathbb{E}_{\mathcal{D}}\left[\left|f - \mathbb{E}[Y \mid f(X), A]\right| \mid A = a\right] . \tag{6}$$

We shall let $\mathbf{suf}_f(A)$ and $\mathbf{cal}_f(A)$ be as defined above in (1) *and* (2)*, respectively.*

2.1. Sufficiency and Calibration

We now state our main results, which show that the sufficiency and calibration gaps of a function f can be controlled by its loss, relative to the calibrated Bayes score f^B. All proofs are deferred to the supplementary material. Throughout, we let $\mathcal{F}$ denote a class of score functions $f : \mathcal{X} \to [0,1]$. For a loss function $\ell : [0,1] \times \{0,1\} \to \mathbb{R}$ and any $\mathcal{D}$-measurable $f : \Omega \to [0,1]$, recall the *population risk* $\mathcal{L}(f) := \mathbb{E}[\ell(f, Y)]$. Note that for $f \in \mathcal{F}$, $\mathcal{L}(f) = \mathbb{E}[\ell(f(X), Y)]$, whereas for the calibrated Bayes score f^B, we denote its population risk as $\mathcal{L}^* := \mathcal{L}(f^B) = \mathbb{E}[\ell(f^B(X, A), Y)]$. We further assume that our losses satisfy the following regularity condition:

Assumption 1. *Given a probability measure $\mathcal{D}$, we assume that $\ell(\cdot,\cdot)$ is (a) κ-strongly convex: $\ell(z, y) \geq \kappa(z - y)^2$, (b) there exists a differentiable map $g : \mathbb{R} \to \mathbb{R}$ such that $\ell(z, y) = g(z) - g(z) - g'(z)(z - y)$ (that is, ℓ is a* Bregman Divergence*), and (c) the calibrated Bayes score is a critical point of the population risk, that is*

$$\mathbb{E}\left[\frac{\partial}{\partial z}\ell(z, Y)\big|_{z=f^B}\right] = 0 .$$

Assumption 1 is satisfied by common choices for the loss function, such as the *square loss* $\ell(z, y) = (z - y)^2$ with $\kappa = 1$, and the logistic loss, as shown by the following lemma, proved in Appendix A.2.

Lemma 2.2 (Logistic Loss)**.** *The logistic loss $\ell(f, Y) = -(Y \log f + (1 - Y) \log(1 - f))$ satisfies Assumption 1 with $\kappa = 2/\log 2$.*

We are now ready to state our main theorem (proved in Appendix B), which provides a simple bound on the sufficiency and calibration gaps, $\mathbf{suf}_f$ and $\mathbf{cal}_f$, in terms of the excess risk $\mathcal{L}(f) - \mathcal{L}^*$:

Theorem 2.3 (Sufficiency and Calibration are Upper Bounded by Excess Risk)**.** *Suppose the loss function $\ell(\cdot,\cdot)$ satisfies Assumption 1 with parameter $\kappa > 0$. Then, for any score $f \in \mathcal{F}$ and any attribute A,*

$$\max\{\mathbf{cal}_f(A), \mathbf{suf}_f(A)\} \leq 4\sqrt{\frac{\mathcal{L}(f) - \mathcal{L}^*}{\kappa}}. \tag{7}$$

Moreover, it holds that for $a \in \mathcal{A}$,

$$\max\{\mathbf{cal}_f(a; A), \mathbf{suf}_f(a; A)\} \leq 2\sqrt{\frac{\mathcal{L}(f) - \mathcal{L}^*}{\Pr[A = a] \cdot \kappa}}. \tag{8}$$

Theorem 2.3 applies to any $f \in \mathcal{F}$, regardless of how f is obtained. As a consequence of Theorem 2.3, we immediately conclude the following corollary for the empirical risk minimizer:

Corollary 2.4 (Calibration of the ERM)**.** *Let $\widehat{f}$ be the output of any learning algorithm (e.g. ERM) trained on a sample $S^n \sim \mathcal{D}^n$, and let $\mathcal{L}(f)$ be as in Theorem 2.3. Then, if $\widehat{f}$ satisfies the guarantee*

$$\Pr_{S^n \sim \mathcal{D}^n}\left[\mathcal{L}(\widehat{f}) - \min_{f \in \mathcal{F}} \mathcal{L}(f) \geq \epsilon\right] \leq \delta,$$

[2] Note that this is *not* the perfect predictor unless Y is deterministic given A and X.

and if ℓ satisfies Assumption 1 with parameter $\kappa > 0$, then with probability at least $1 - \delta$ over $S^n \sim \mathcal{D}^n$, it holds that

$$\max\{\mathbf{cal}_f(A), \mathbf{suf}_f(A)\} \leq 4\sqrt{\frac{\epsilon + \min_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}^*}{\kappa}}.$$

The above corollary states that if there exists a score in the function class $\mathcal{F}$ whose population risk $\mathcal{L}(f)$ is close to that of the calibrated Bayes optimal $\mathcal{L}^*$, then empirical risk minimization succeeds in finding a well-calibrated score.

In order to apply Corollary 2.4, one must know when the gap between the best-in-class risk and calibrated Bayes risk, $\min_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}^*$, is small. In the full information setting where $A = A(X)$ (that is, the group attribute is available to the score function), $\min_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}^*$ corresponds to the approximation error for the class $\mathcal{F}$ (Bartlett et al., 2006). When X may not contain all the information about A, $\min_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}^*$ depends not only on the class $\mathcal{F}$ but also on how well A can be encoded by X given the class $\mathcal{F}$, and possibly additional regularity conditions. We now present a guiding example under which one can meaningfully bound the excess risk in the incomplete information setting. In Appendix B.3, we provide two further examples to guide the readers' intuition. For our present example, we introduce as a benchmark the *uncalibrated Bayes optimal score*

$$f^U(x) := \mathbb{E}[Y|X = x],$$

which minimizes empirical risk over all X measurable functions, and is necessarily in $\mathcal{F}$. Our first example gives a decomposition of $\mathcal{L}(f) - \mathcal{L}^*$ when ℓ is the square loss.

Example 2.1. *Let $\ell(z, y) := (z - y)^2$ denote the squared loss. Then,*

$$\mathcal{L}(\widehat{f}) - \mathcal{L}^* = \underbrace{\left(\mathcal{L}(\widehat{f}) - \inf_{f \in \mathcal{F}}(f)\right)}_{(i)} + \underbrace{\left(\inf_{f \in \mathcal{F}} \mathcal{L}(f) - \mathcal{L}(f^U)\right)}_{(ii)} + \underbrace{\mathbb{E}_X\left[\mathrm{Var}_A\left[f^B \mid X\right]\right]}_{(iii)}, \quad (9)$$

where $\mathrm{Var}_A[f^B \mid X] = \mathbb{E}[(f^B - \mathbb{E}_A[f^B \mid X])^2 \mid X]$ denotes the conditional variance of f^B given X.

The decomposition in Example 2.1 follows immediately from the fact that the excess risk of f^U over f^B, $\mathcal{L}(f^U) - \mathcal{L}^*$, is precisely $\mathrm{Var}_A[f^B \mid X]$ when ℓ is the square loss. Examining (9), (i) represents the excess risk of $\widehat{f}$ over the best score in $\mathcal{F}$, which tends to zero if $\widehat{f}$ is the ERM. Term (ii) captures the richness of the function class, for as $\mathcal{F}$ contains a close approximation to f^U. If $\widehat{f}$ is obtained by a consistent non-parametric learning procedure, and f^U has small complexity, then both (i) and (ii) tend to zero in the limit of infinite samples. Lastly, (iii) captures the additional information about A contained in X. Note that in the full information zero, this term is zero.

2.2. Lower Bounds for Separation

In this section, we show that empirical risk minimization robustly violates the *separation* criterion that scores are conditionally independent of the group A given the outcome Y. For a classifier that exactly satisfies separation, we have $\mathbb{E}[f(X) \mid Y, A] = \mathbb{E}[f(X) \mid Y]$ for any group A and outcome Y. We define the *separation gap* as the average margin by which this equality is violated:

Definition 2 (Separation gap)**.** *The* separation gap *is*

$$\mathbf{sep}_f(A) := \mathbb{E}_{Y,A}[|\mathbb{E}[f(X) \mid Y, A] - \mathbb{E}[f(X) \mid Y]|].$$

Our first result states that the calibrated Bayes score f^B, has a non-trivial separation gap. The following lower bound is proved in Appendix F:

Proposition 2.5 (Lower bound on separation gap)**.** *Denote $\overline{q} := \Pr[Y = 1]$, and $q_A := \Pr[Y = 1|A]$ for a group attribute A. Let $\mathrm{Var}(\cdot)$ denote variance, and $\mathrm{Var}(\cdot \mid X)$ denote conditional variance given a random variable X. Then, $\mathbf{sep}_{f^B}(A) \geq C_{f^B} \cdot Q_A$, where*

$$Q_A := \mathbb{E}_A|\overline{q} - q_A| \quad \textit{and} \quad C_{f^B} := \frac{\mathbb{E}_{\mathcal{D}}\mathrm{Var}[Y \mid X, A]}{\mathrm{Var}[Y]}.$$

Intuitively, the above bound says that the separation gap of the calibrated Bayes score is lower bounded by the product of two quantities: $Q_A = \mathbb{E}_A|q_A - \overline{q}|$ corresponds to the L_1-variation in base-rates among groups, and C_{f^B} corresponds to the intrinsic noise level of the prediction problem. For example, consider the case where perfect prediction is possible (that is, Y is deterministic given X, A). Then, the lower bound is vacuous because $C_f^B = 0$, and indeed f^B has zero separation gap.

Proposition 2.5 readily implies that any score f which has small risk with respect to f^B also necessarily violates the separation criterion:

Corollary 2.6 (Separation of the ERM)**.** *Let $\mathcal{L}$ be the risk associated with a loss function $\ell(\cdot, \cdot)$ satisfying Assumption 1 with parameter $\kappa > 0$. Then, for any score $\hat{f} \in \mathcal{F}$, possibly the ERM, and any attribute A,*

$$\mathbf{sep}_{\hat{f}} \geq C_{f^B} \cdot \mathbb{E}_A|q_A - \overline{q}| - 2\sqrt{\frac{\mathcal{L}(\hat{f}) - \mathcal{L}^*}{\kappa}}.$$

In prior work, Kleinberg et al. (2017)'s impossibility result (Theorem 1.1, 1.2), as well as subsequent generalizations in Pleiss et al. (2017), states that a score that satisfies both calibration and separation must be either a perfect predictor or the problem must have equal base rates across groups, that is, $\overline{q} = q_A$. In contrast, Proposition 2.5 provides a quantitative lower bound on the separation gap of a calibrated score, for arbitrary configurations of base rates and closeness to perfect prediction. This is crucial for approximating the separation gap of the ERM in Corollary 2.6.

2.3. Lower Bounds for Sufficiency and Calibration

We now present two lower bounds which demonstrate that the behavior depicted in Theorem 2.3 is sharp in the worse case. In Appendix C, we construct a family of distributions $\{\mathcal{D}_\theta\}_{\theta\in\Theta}$ over pairs $(X, Y) \in \mathcal{X} \times \{0, 1\}$, and a family of attributes $\{A_w\}_{w\in\mathcal{W}}$ which are measurable functions of X. We choose the distribution parameter θ and attribute parameter w to be drawn from specified priors π_Θ and $\pi_\mathcal{W}$. We also consider a class of score functions $\mathcal{F}$ mapping $\mathcal{X} \to [0, 1]$, which contains the calibrated Bayes classifer for any $\theta \in \Theta$ and $w \in \mathcal{W}$ (this is possible because the attributes are X-measurable). We choose $\mathcal{L}$ to be the risk associated with the square loss, and consider classifiers trained on a sample $S^n = \{(X_i, Y_i)\}_{i=1}^n$ of n i.i.d draws from $\mathcal{D}_\theta$. In this setting, we have the following:

Theorem 2.7. *Let $\widehat{f} \in \mathcal{F}$ denote the output of any learning algorithm trained on a sample $S^n \sim \mathcal{D}^n$, and let $\widehat{f}_n$ denote the empirical risk minimizer of $\mathcal{L}$ trained on S^n. Then, with constant probability over $\theta \sim \pi_\Theta$, $w \sim \pi_\mathcal{W}$, and $S^n \sim \mathcal{D}_\theta$, $\min\{\mathbf{cal}_{\widehat{f}}(A_w), \mathbf{suf}_{\widehat{f}}(A_w)\} \geq \Omega(1/\sqrt{n})$ and $\mathcal{L}(\widehat{f}_n) - \mathcal{L}* \leq O(1/n)$.*

In particular, taking $\widehat{f} = \widehat{f}_n$, we see that the for any sample size n, we have that

$$\min\{\mathbf{cal}_{\widehat{f}_n}(A_w), \mathbf{suf}_{\widehat{f}_n}(A_w)\}/\sqrt{\mathcal{L}(\widehat{f}_n) - \mathcal{L}*} = \Omega(1).$$

with constant probability. In addition, Theorem 2.7 shows that in the worst case, the calibration and sufficiency gaps decay as $\Omega(1/\sqrt{n})$ with n samples.

We can further modify the construction to lower bound the per-group sufficency and calibration gaps in terms of $\Pr[A = a]$. Specifically, for each $p \in (0, 1/4)$, we construct in Appendix D a family of distributions $\{\mathcal{D}_{\theta;p}\}_{\theta\in\Theta}$ and X-measurable attributes $\{A_w\}_{w\in\mathcal{W}}$ such that, for all (θ, w), $\min_{a\in\mathcal{A}} \Pr_{(X,Y)\sim\mathcal{D}_{\theta;p}}[A_w(X) = a] = p$, for all $\theta \in \Theta$ and $w \in \mathcal{W}$. The construction also entails modifying the class $\mathcal{F}$; in this setting, our construction is as follows:

Theorem 2.8. *Fix $p \in (0, 1/4)$. For any score $\widehat{f} \in \mathcal{F}$ trained on S^n, and the empirical risk mnimizer $\widehat{f}_n$, it holds that $\min\{\mathbf{cal}_{\widehat{f}}(A_w), \mathbf{suf}_{\widehat{f}}(A_w)\} \geq \Omega(1/\sqrt{pn})$ and $\mathcal{L}(\widehat{f}_n) - \mathcal{L}* \leq O(1/n)$, with constant probability over $\theta \sim \pi_\Theta$, $w \sim \pi_\mathcal{W}$, and $S^n \sim \mathcal{D}_{\theta;p}$.*

3. Experiments

In this section, we present numerical experiments on two datasets to corroborate our theoretical findings. These are the Adult dataset from the UCI Machine Learning Repository (Dua and Karra Taniskidou, 2017) and a dataset of pretrial defendants from Broward County, Florida (Angwin et al., 2016; Dressel and Farid, 2018) (henceforth referred to as the Broward dataset).

The Adult dataset contains 14 demographic features for 48842 individuals, for predicting whether one's annual income is greater than $50,000. The Broward dataset contains 7 features of 7214 individuals arrested in Broward County, Florida between 2013 and 2014, with the goal of predicting recidivism within two years. It is derived by Dressel and Farid (2018) from the original dataset used by Angwin et al. (2016) to evaluate a widely used criminal risk assessment tool. We present results for the Adult dataset in the current section, and those for the Broward dataset in Appendix G.2.

Score functions are obtained by logistic regression on a training set that is 80% of the original dataset, using all available features, unless otherwise stated.

We first examine the sufficiency of the score with respect to two sensitive attributes, gender and race in Section 3.1. Then, in Section 3.2 we show that the score obtained from empirical risk minimization is sufficient and calibrated with respect to multiple sensitive attributes simultaneously. Section 3.3 explores how sufficiency and separation are affected differently by the amount of training data, as well as the model class.

We use two descriptions of sufficiency. In Sections 3.1 and 3.2, we present the so-called *calibration plots* (e.g., Figure 1), which plots observed positive outcome rates against score deciles for different groups. The shaded regions indicate 95% confidence intervals for the rate of positive outcomes under a binomial model. In Section 3.3, we report empirical estimates of the sufficiency gap, $\mathbf{suf}_f(A)$, using a test set that is 20% of the original dataset. More details on this estimator can be found in Appendix G.1. In general, models that are more sufficient and calibrated have smaller $\mathbf{suf}_f$ and their calibration plots show overlapping confidence intervals for different groups.

3.1. Training with Group Information has Modest Effects on Sufficiency

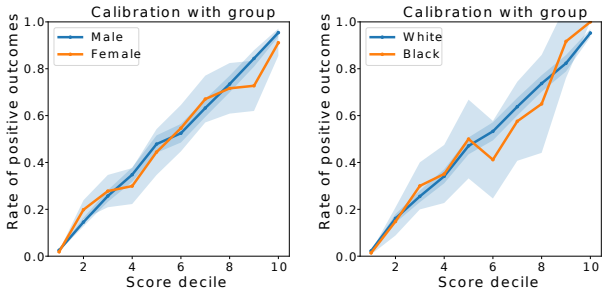

Figure 1. Calibration plot for score using group attribute

In this section, we examine the sufficiency of ERM scores, with respect to gender and race. When all available features were used in the regression, including sensitive attributes,

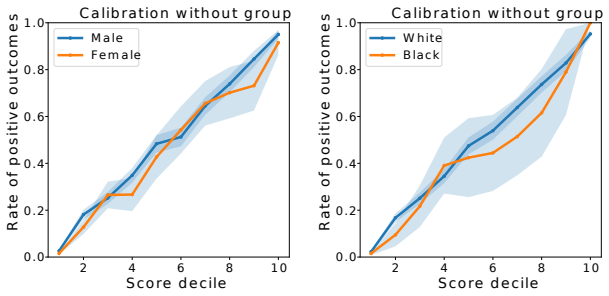

Figure 2. Calibration plot for score not using group attribute

the empirical risk minimizer of the logistic loss is sufficient and calibrated with respect to both gender and race, as seen in Figure 1. However, sufficiency can hold approximately even when the score is not a function of the group attribute. Figure 2 shows that without the group variable, the ERM score is only slightly less calibrated; the confidence intervals for both groups still overlap at every score decile.

3.2. Simultaneous Sufficiency with respect to Multiple Group Attributes

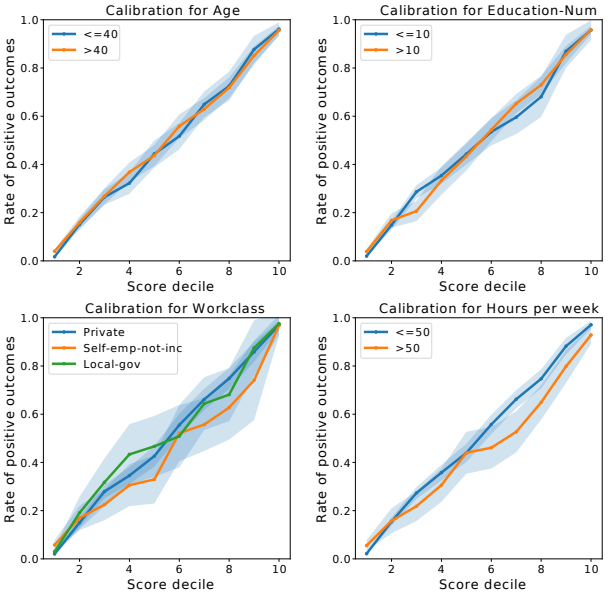

Figure 3. Calibration plot with respect to other group attributes

Furthermore, we observe that empirical risk minimization with logistic regression also achieves approximate sufficiency with respect to any other group attribute defined on the basis of the given features, not only gender and race. In Figure 3, we show the calibration plot for the ERM score with respect to Age, Education-Num, Workclass, and Hours per week; Figure 4 considers combinations of two features. In each case, the confidence intervals for the rate of positive outcomes for all groups overlap at all, if not most, score

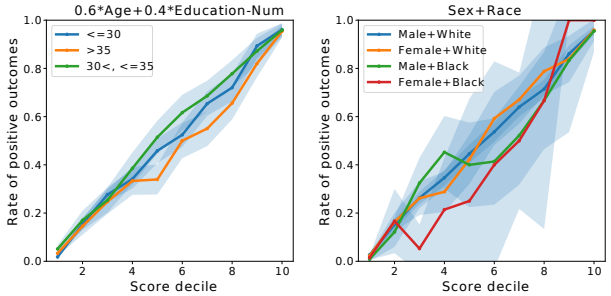

Figure 4. Calibration plot with respect to combinations of features: linear combination (left), intersectional combination (right)

deciles. In particular, Figure 4 (right) shows that the ERM score is close to sufficient and calibrated even for a newly defined group attribute that is the intersectional combination of race and gender. The calibration plots for other features, as well as implementation details, can be found in Appendix G.3.

3.3. Sufficiency Improves with Model Accuracy and Model Flexibility

Our theoretical results suggest that the sufficiency gap of a score function is tightly related to its excess risk. In general, it is impossible to determine the excess risk of a given classifier with respect to the Bayes risk $\mathcal{L}^*$ from experimental data. Instead we shall examine how the sufficiency gap of a score trained by logistic regression varies with the number of samples and the model class, both of which were chosen because of their impact on the excess risk of the score.

Specifically, we explore the effects of decreased risk on sufficiency gap due to (a) increased number of training examples (Figure 5) and (b) increased expressiveness of the class $\mathcal{F}$ of score functions (Figure 6). As the number of training samples increases, the gap between the ERM and least-risk score function in a given class $\mathcal{F}$, $\operatorname{argmin}_{f\in\mathcal{F}} \mathcal{L}(f)$, decreases. On the other hand, as the number of model parameters grows, the class $\mathcal{F}$ becomes more expressive, and $\min_{f\in\mathcal{F}} \mathcal{L}(f)$ may become closer to the Bayes risk $\mathcal{L}^*$.

Figures 5 and 6 display, for each experiment, the sufficiency gap and logistic loss on a test set averaged over 10 random trials, each using a randomly chosen training set. The shaded region in the figures indicates two standard deviations from the average value. In Figure 5, as the number of training examples increase, the logistic loss of the score decreases, and so does the sufficiency gap. For the race group attribute, we even observe that the sufficiency gap is going to zero; this is predicted by Theorem 2.3 as the risk of the score approaches the Bayes risk. Figure 5 also displays the separation gap of the scores. Indeed, the separation gap is bounded away from zero, as predicted by Corollary 2.6,

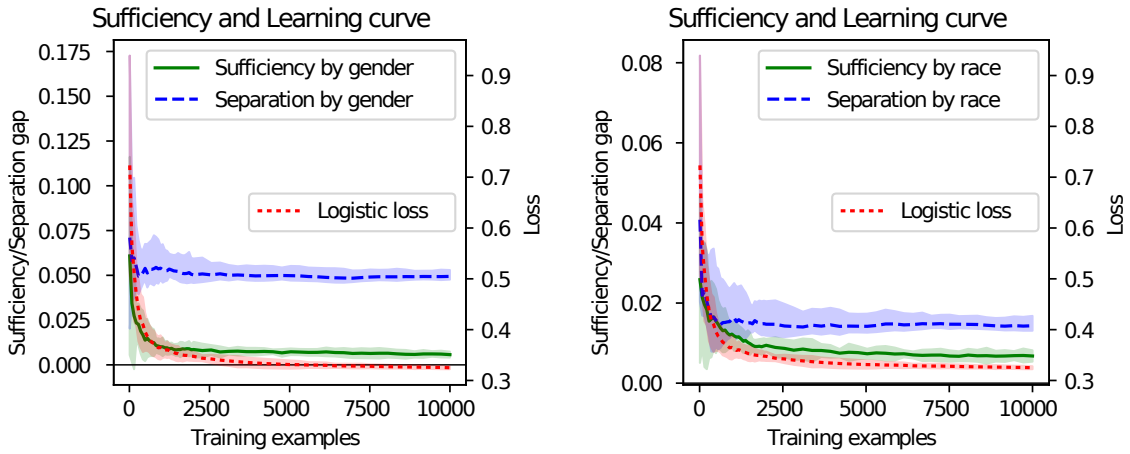

Figure 5. Sufficiency, Separation, and Logistic Loss vs. Number of training examples

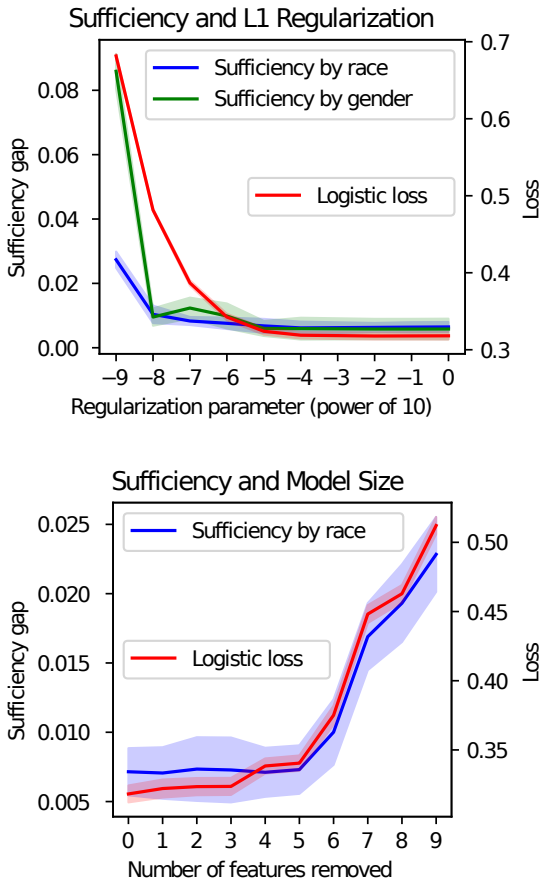

Figure 6. Sufficiency for models trained with different L1 regularization parameters (top) and with different number of features (bottom)

and does not decrease with the number of training examples. This corroborates our finding that unconstrained machine learning cannot achieve the separation notion of fairness even with infinite data samples.

In Figure 6 (bottom), we gradually restrict the model class by reducing the number of features used in logistic regression. As the number of features decreases, the logistic loss increases and so does the sufficiency gap. In Figure 6 (top), we implicitly restrict the model class by varying the regularization parameter: with a smaller parameters corresponding to more severe regularization, constraining the learned weights to be inside a smaller L1 ball. As we increase regularization, the logistic loss increases and so does the sufficiency gap. Both experiments show that the sufficiency gap is reduced when the model class is enlarged, again demonstrating its tight connection to the excess risk.

Conclusion Our results show that group calibration follows from closeness to the risk of the calibrated Bayes optimal score function. Consequently, unconstrained machine learning (e.g. via empirical risk minimization or some surrogate thereof) may serve as a simple recipe for achieving group calibration, provided that (1) the function class is sufficiently rich, (2) there are enough training samples, and (3) the group attribute can be approximately predicted from the available features. On the other hand, we show that group calibration does not and cannot solve fairness concerns that pertain to the Bayes optimal score function, such as the violation of separation and independence.

Our findings suggest that group calibration is an appropriate notion of fairness *only* when we expect unconstrained machine learning to be fair, given sufficient data. Thus, focusing on calibration alone is likely insufficient to mitigate the negative impacts of unconstrained machine learning.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *UCLA Law Review*, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. http://www.fairmlbook.org.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5, 2017.

Alexandra Chouldechova and Aaron Roth. The Frontiers of Fairness in Machine Learning. *CoRR*, abs/1810.08810, 2018.

T. Anne Cleary. Test bias: Validity of the scholastic aptitude test for negro and white students in integrated colleges. *ETS Research Bulletin Series*, 1966(2):i–23.

T. Anne Cleary. Test bias: Prediction of grades of negro and white students in integrated colleges. *Journal of Educational Measurement*, 5(2):115–124, 1968.

Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *CoRR*, abs/1808.00023, 2018.

Sam Corbett-Davies, Sharad Goel, and Sandra Gonzlez-Bailn. Thoughts on machine learning accuracy. *New York Times*, July 2017a.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pages 797–806, New York, NY, USA, 2017b. ACM.

David R. Cox. Two further applications of a model for binary regression. *Biometrika*, 45(3-4):562–565, 1958.

Kate Crawford. The hidden biases in big data. *Harvard Business Review*, 1, 2013.

Kate Crawford. The trouble with bias. NeurIPS Keynote , 2017.

Richard B Darlington. Another look at "cultural fairness". *Journal of Educational Measurement*, 8(2):71–82, 1971.

A. P. Dawid. The well-calibrated bayesian. *Journal of the American Statistical Association*, 77(379):605–610, 1982.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(1/2): 12–22, 1983.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity, 2016.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1), 2018. doi: 10.1126/sciadv.aao5580.

Dheeru Dua and Efi Karra Taniskidou. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, 2015.

M. Hardt, E. Price, and N. Srebo. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*, 2018.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1944–1953, Stockholm, Sweden, 2018.

Michael J. Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *CoRR*, abs/1711.05144, 2017.

Jon Kleinberg, Jens Ludwig, and Sendhil Mullainathan. A guide to solving social problems with machine learning. *Harvard Business Review*, December 2016.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *Proc. 8th ITCS*, 2017.

Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3156–3164, Stockholm, Sweden, 2018.

Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):41–47, 1977.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, pages 625–632, 2005.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 1999.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems 30*, pages 5684–5693, 2017.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P. Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 2239–2248, New York, NY, USA, 2018. ACM.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, pages 831–838, 1992.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 609–616, 2001.