The implicit fairness criterion of unconstrained learning

Lydia T. Liu*† Max Simchowitz*† Moritz Hardt*

Abstract

We clarify what fairness guarantees we can and cannot expect to follow from unconstrained machine learning. Specifically, we characterize when unconstrained learning on its own implies *group calibration*, that is, the outcome variable is conditionally independent of group membership given the score. We show that under reasonable conditions, the deviation from satisfying group calibration is upper bounded by the excess risk of the learned score relative to the Bayes optimal score function. 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 show that group calibration is the fairness criterion that unconstrained learning implicitly favors. On the one hand, this means that calibration is often satisfied on its own without the need for active intervention, albeit at the cost of violating other criteria that are at odds with calibration. On the other hand, it suggests that we should be satisfied with calibration as a fairness criterion only if we are at ease with the use of unconstrained machine learning in a given application.

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 simply seeks to minimize the average prediction loss over a labeled dataset, 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 [Corbett-Davies et al., 2017b,a, Kleinberg et al., 2018].

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

*Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
†Equal contribution

detention. After journalists at ProPublica pointed out that a popular recidivism risk score 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]. 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].

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 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 Our 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 particular, our results *do not* require that X perfectly encodes the attribute A.

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*:

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

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

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

[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.

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.

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

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

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. Such impossibility results have been further generalized 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 fairness, such as the impact of model predictions on future welfare [Liu et al., 2018] and demographics [Hashimoto et al., 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 as*

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

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

and the calibration gap for group a as

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

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*

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

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,*

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

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

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, the empirical risk

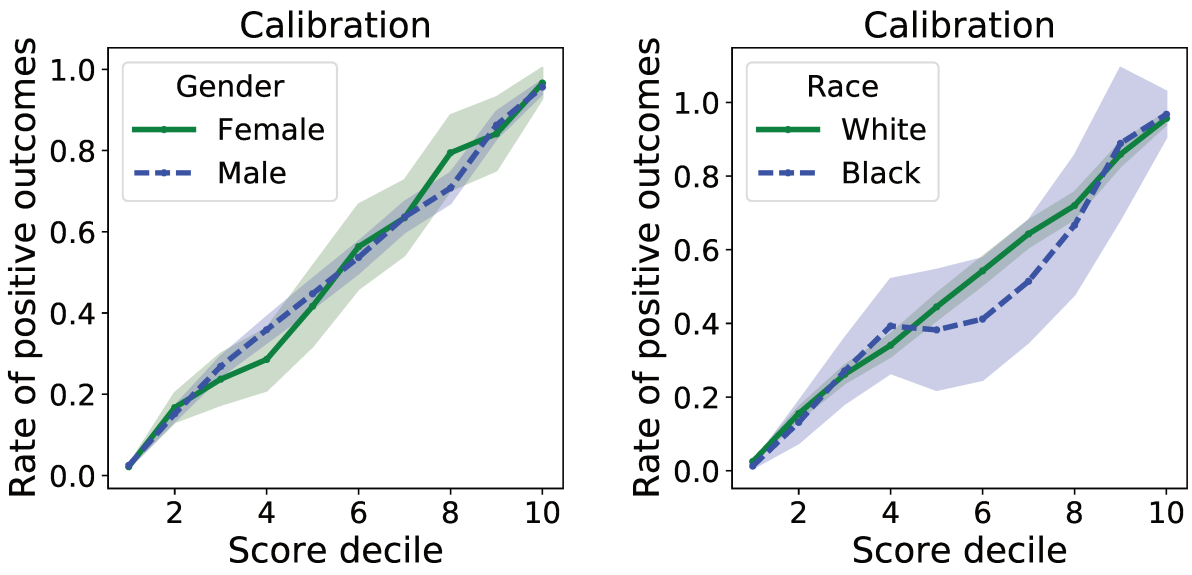

Figure 1: Calibration plot for score using group attribute

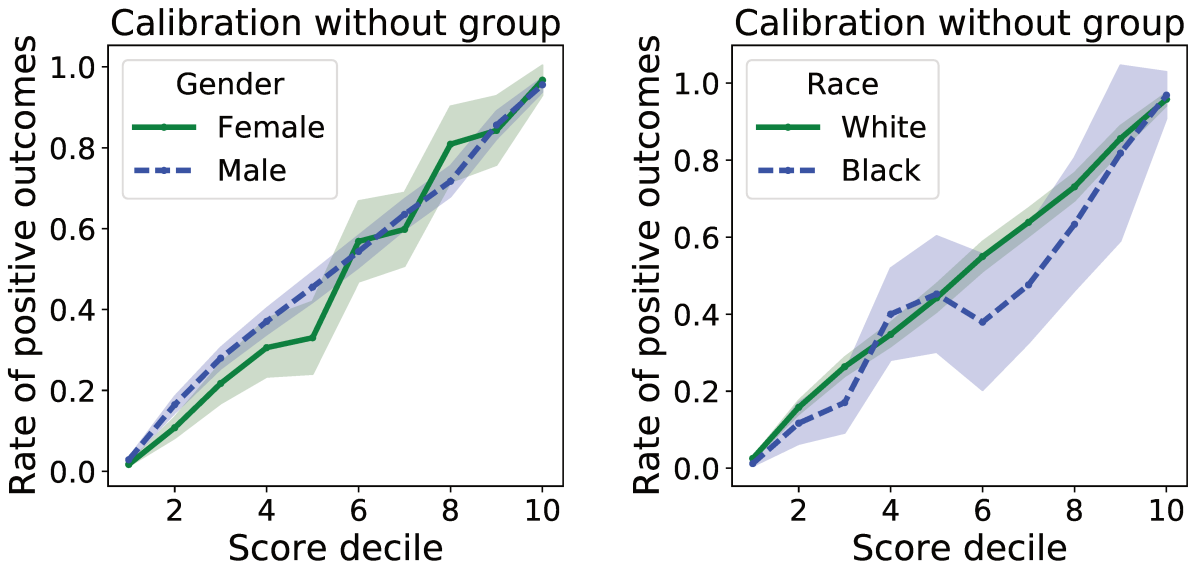

Figure 2: Calibration plot for score not using group attribute

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

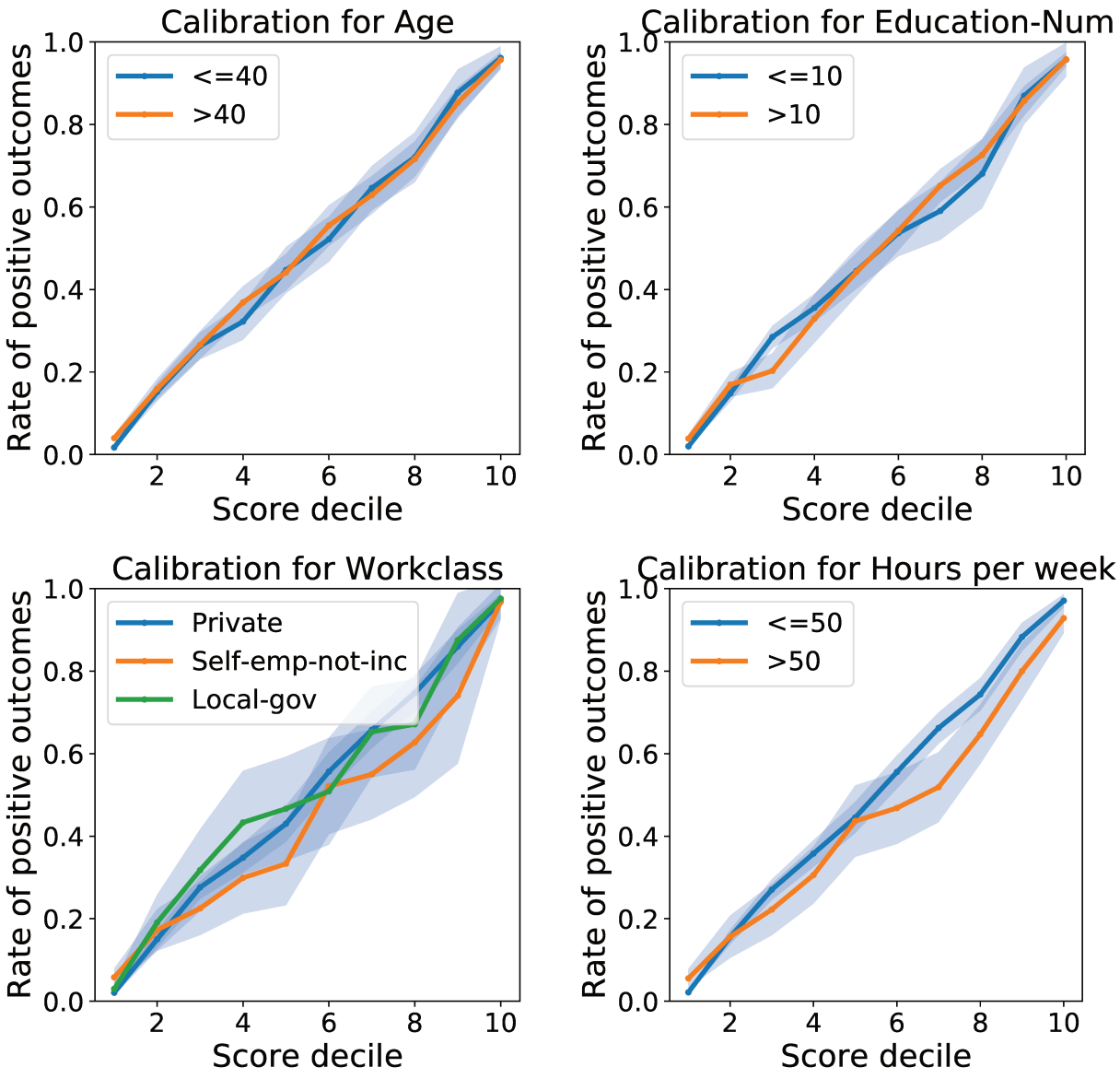

Figure 3: Calibration plot with respect to other group attributes

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

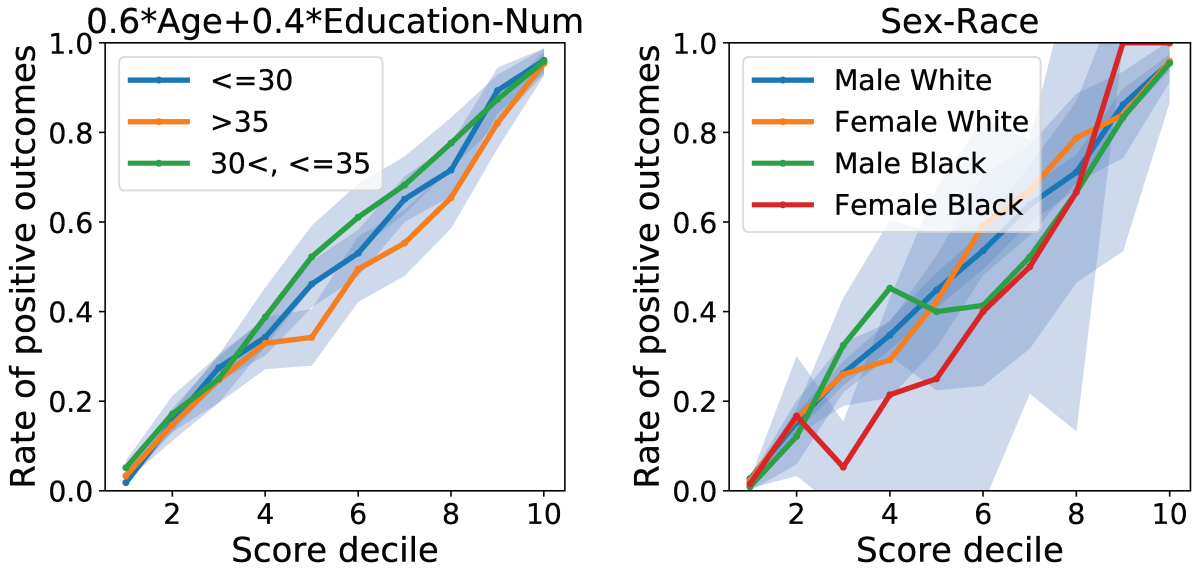

Figure 4: Calibration plot with respect to combinations of features: linear combination (left), intersectional combination (right)

overlap at all, if not most, score 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

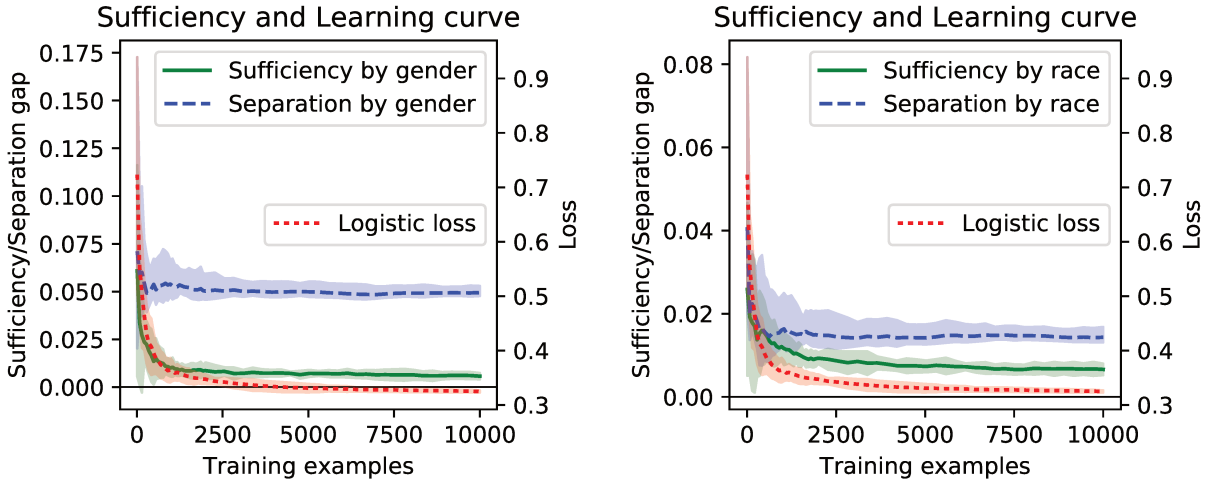

Figure 5: Sufficiency, Separation, and Logistic Loss vs. Number of training examples

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

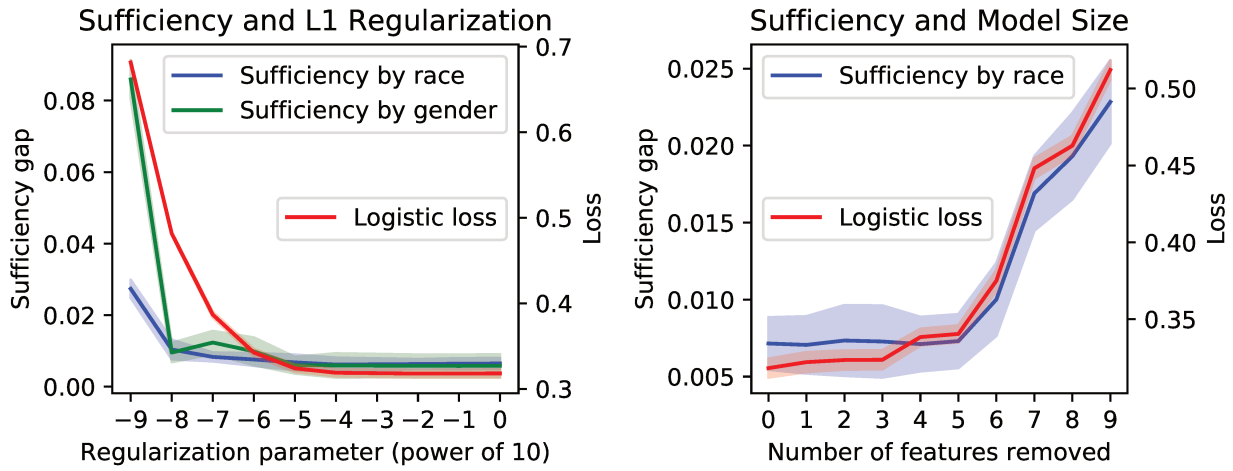

Figure 6: Sufficiency for models trained with different L1 regularization parameters (left) and with different number of features (right)

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, 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 (right), 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 (left), we implicitly restrict the model class by varying the regularization parameter. In this case, a smaller regularization parameter corresponds to more severe regularization, which constrains 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 In summary, our results show that group calibration follows from closeness to the risk of the calibrated Bayes optimal score function. Consequently, empirical risk minimization is a simple and efficient 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.

More broadly, 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. Stated otherwise, focusing on calibration alone is likely insufficient to mitigate the negative impacts of unconstrained machine learning.

References

Rudolf Ahlswede. The final form of Tao's inequality relating conditional expectation and conditional mutual information. *Advances in Mathematics of Communications*, 1:239, 2007.

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.

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, Sharad Goel, and Sandra Gonzlez-Bailn. Thoughts on machine learning accuracy. *New York Times*, July 2017a. URL https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html.

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. NIPS Keynote https://www.youtube.com/watch?v=fMym_BKWQzk, 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. URL https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html.

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.

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 (NIPS)*, 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.

Daniel Hsu, Sham M Kakade, and Tong Zhang. An analysis of random design linear regression. Citeseer, 2011.

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, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. *AEA Papers and Proceedings*, 108:22–27, 2018.

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. ISBN 1-59593-180-5.

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.

Max Simchowitz, Kevin Jamieson, and Benjamin Recht. Best-of-K-bandits. In *Conference on Learning Theory*, pages 1440–1489, 2016.

Terence Tao. Szemerédi's regularity lemma revisited. *Contributions to Discrete Mathematics*, 1, 2006.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519, 9780387790510.

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. ISBN 1-55860-778-1.

A Additional Proofs for Section 2.1

In this section we prove Proposition 2.1, and Lemma 2.2.

A.1 Proof of Proposition 2.1

Recall that $f^B(X, A) = \mathbb{E}[Y \mid X, A]$.

By the tower rule for conditional expectation,

$$\begin{aligned}\Pr\left[Y = 1 \mid f^B(X, A), A\right] &= \mathbb{E}[Y \mid f^B(X, A), A] \\ &= \mathbb{E}\left[\mathbb{E}[Y \mid X, A] \mid f^B(X, A), A\right] \\ &= \mathbb{E}[f^B(X, A) \mid f^B(X, A), A] \;=\; f^B(X, A)\ ,\end{aligned}$$

and,

$$\begin{aligned}\Pr[Y = 1 \mid f^B(X, A)] &= \mathbb{E}[Y \mid f^B(X, A)] \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, A] \mid f^B(X, A)] \\ &= \mathbb{E}[f^B(X, A) \mid f^B(X, A)] = f^B(X, A).\end{aligned}$$

Therefore, the calibrated Bayes score $f^B(X)$ is sufficient and calibrated.

More generally, conditional expectations of the form $f(X) = \mathbb{E}[Y \mid \Phi(X), A]$ are calibrated, where $\Phi : \mathcal{X} \to \mathcal{X}'$ can be any transformation of the features. This follows similarly from the tower rule.

A.2 Proof of Lemma 2.2

To see that this is true, first note that $\ell(f, y)$ is a Bregman divergence. We can easily check that $\mathbb{E}[\nabla_f \ell(f(x, a), Y)|_{f^B}] = \mathbb{E}[\frac{Y}{f^B} - \frac{1-Y}{1-f^B}] = 0$. Finally, κ-strong convexity follows from Pinkser's inequality for Bernoulli random variables:

$$(f - f')^2 \leq \frac{\log 2}{2}\left(f' \log \frac{f'}{f} + (1 - f') \log \frac{1 - f'}{1 - f}\right) = \frac{\log 2}{2}\ell(f, f').$$

B Proof of Theorem 2.3

Throughout, we consider a fixed distribution $\mathcal{D}$ and attribute A. We shall also use the shorthand $f = f(X)$ and $f^B = f^B(X, A)$. We begin by proving the following lemma, which establishes Theorem 2.3 in the case where f is the squared loss:

Lemma B.1. *Let f^B be the Bayes classifier, and let f denote any function. Then,*

$$\mathbf{suf}_f \leq 4\sqrt{\mathbb{E}_{X,A}[(f - f^B)^2]}, \tag{10}$$

$$\forall a \in \mathcal{A}, \quad \mathbf{suf}_f(a; A) \leq 2\sqrt{\frac{\mathbb{E}_{X,A}[(f - f^B)^2]}{\Pr[A = a]}}. \tag{11}$$

$$\mathbf{cal}_f \leq \sqrt{\mathbb{E}_{X,A}[(f - f^B)^2]}, \tag{12}$$

$$\forall a \in \mathcal{A}, \quad \mathbf{cal}_f(a; A) \leq 2\sqrt{\frac{\mathbb{E}_{X,A}[(f - f^B)^2]}{\Pr[A = a]}}. \tag{13}$$

To conclude the proof of Theorem 2.3, we first show that $\mathbb{E}[\ell(f, f^B)] = \mathbb{E}[\ell(f, Y)] - \mathbb{E}[\ell(f^B, Y)]$. Since ℓ is a Bregman divergence and calibrated at f^B (Assumption 1), we have

$$\begin{aligned}\mathbb{E}[\ell(f, f^B)] =& \mathbb{E}[\ell(f, Y)] - \mathbb{E}[\ell(f^B, Y)] + \mathbb{E}[(g'(Y) - g'(f^B)) \cdot (f - f^B)] \\ =& \mathbb{E}[\ell(f, Y)] - \mathbb{E}[\ell(f^B, Y)] - \mathbb{E}[\underbrace{\mathbb{E}[\nabla_f \ell(f, Y)|_{f^B} \mid X, A]}_{=0} \cdot (f - f^B)] \\ =& \mathbb{E}[\ell(f, Y)] - \mathbb{E}[\ell(f^B, Y)].\end{aligned}$$

Moreover, by strong convexity, we have that $\mathcal{L}(f) \geq \frac{1}{\kappa}\mathbb{E}[(f - Y)^2]$. Thus,

$$\kappa\mathbb{E}[(f - f^B)^2] \leq \mathbb{E}[\ell(f, f^B)] = \mathbb{E}[\ell(f, Y)] - \mathbb{E}[\ell(f^B, Y)] = \mathcal{L}(f) - L(f^B) \ .$$

Applying Lemma B.1 concludes the proof.

B.1 Proof of Lemma B.1 (10) and (11)

First we bound the $\mathcal{L}_2$ difference of the conditional expectations. Note that since $f = f(X)$,

$$\mathbb{E}[Y \mid f, A] = \mathbb{E}[\mathbb{E}[Y|X, A, f] \mid f, A] = \mathbb{E}[\mathbb{E}[Y|X, A] \mid f, A] = \mathbb{E}[f^B \mid f, A]. \tag{14}$$

Moreover, by the definition of f^B

$$\begin{aligned}\mathbb{E}[Y \mid f^B, A] = \mathbb{E}[\mathbb{E}[Y \mid A, X, f^B], f^B, A] \ &= \ \mathbb{E}[\mathbb{E}[Y \mid A, X, \mathbb{E}[Y \mid A, X]] \\ = \mathbb{E}[\mathbb{E}[Y \mid A, X], A, X] = \mathbb{E}[Y \mid A, X] &= f^B,\end{aligned} \tag{15}$$

and thus, by (14) and (15), we have

$$\begin{aligned}\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f, A] - \mathbb{E}[Y \mid f^B, A])^2] &= \mathbb{E}_{X,A}[(\mathbb{E}[f^B \mid f, A] - f^B)^2] \quad \text{by (14) and (15)} \\ &= \mathbb{E}_{X,A}[(\mathbb{E}[f^B - f \mid f, A] + f - f^B)^2] \\ &\leq 2\mathbb{E}_{X,A}[(\mathbb{E}[f^B - f \mid f, A])^2 + (f - f^B)^2] \\ &\leq 2\mathbb{E}_{X,A}[\mathbb{E}[(f^B - f)^2 \mid f, A] + (f - f^B)^2] \qquad (16) \\ &= 4\mathbb{E}_{X,A}[(f - f^B)^2] \ , \qquad (17)\end{aligned}$$

where (16) follows from Jensen's inequality. Similarly, one has

$$\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f^B])^2] \leq 4\mathbb{E}_{X,A}[(f - f^B)^2]. \tag{18}$$

We then find that

$$\begin{aligned}\Delta_f &= \Delta_f - \Delta_{f^B} \\ &= \mathbb{E}_{X,A}[|\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A]| - |\mathbb{E}[Y \mid f^B] - \mathbb{E}[Y \mid f^B, A]|] \\ &= \mathbb{E}_{X,A}[|\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A]| - |\mathbb{E}[Y \mid f^B] - \mathbb{E}[Y \mid f^B, A]|] \\ &= \mathbb{E}_{X,A}[\sqrt{(\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A] - (\mathbb{E}[Y \mid f^B] - \mathbb{E}[Y \mid f^B, A]))^2}] \\ &\leq \sqrt{\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A] - (\mathbb{E}[Y \mid f^B] - \mathbb{E}[Y \mid f^B, A]))^2]} \qquad (19) \\ &\leq \sqrt{2\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f^B])^2] + 2\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f, A] - \mathbb{E}[Y \mid f^B, A])^2]} \\ &\leq \sqrt{8\mathbb{E}_{X,A}[(f - f^B)^2] + 8\mathbb{E}_{X,A}[(f - f^B)^2]} \ = 4\sqrt{\mathbb{E}_{X,A}[(f - f^B)^2]}, \qquad (20)\end{aligned}$$

where we've applied Jensen's inequality in (19), and the inequality in (20) uses (17) and (18).

Similarly, for a fixed group $A = a$, we have

$$\begin{aligned}\mathbf{suf}_f(a; A) &= \mathbb{E}_X[|\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A]| - |\mathbb{E}[Y \mid f^B] - \mathbb{E}[Y \mid f^B, A]| \mid A = a] \\ &\leq 4\sqrt{\mathbb{E}_X[(f - f^B)^2 \mid A = a]} \\ &\leq 4\sqrt{\frac{1}{\Pr[A = a]}\mathbb{E}_{X,A}[(f - f^B)^2]}\end{aligned}$$

B.2 Proof of Lemma B.1 (12) and (13)

By (14), we have

$$\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f, A] - f)^2] = \mathbb{E}_{X,A}[(\mathbb{E}[f^B \mid f, A] - f)^2] \leq \mathbb{E}_{X,A}[(f^B - f)^2], \tag{21}$$

where the inequality follows from Jensen's inequality and the tower property. We then find that, by Jensen's inequality,

$$\begin{aligned}\mathbf{cal}_f &= \mathbb{E}_{X,A}[|\mathbb{E}[Y \mid f, A] - f|] \\ &= \mathbb{E}_{X,A}[\sqrt{(\mathbb{E}[Y \mid f, A] - f)^2}] \\ &\leq \sqrt{\mathbb{E}_{X,A}[(\mathbb{E}[Y \mid f, A] - f)^2]} \\ &\leq \sqrt{\mathbb{E}_{X,A}[(f^B - f)^2]}.\end{aligned}$$

Similarly, for an fixed group $A = a$, we have

$$\begin{aligned}\mathbf{cal}_f(a; A) &= \mathbb{E}_{X,A}[|\mathbb{E}[Y \mid f, A] - f| \mid A = a] \\ &\leq \sqrt{\mathbb{E}_X[(f - f^B)^2 \mid A = a]} \\ &\leq \sqrt{\frac{1}{\Pr[A = a]}\mathbb{E}_{X,A}[(f - f^B)^2]}.\end{aligned}$$

B.3 Further Examples for Calibration and Sufficiency Bounds

We present two further examples under which one can meaningfully bound the excess risk, and consequently the sufficiency and calibration gaps, in the incomplete information setting. In the next example, we examine sufficiency when $\widehat{f}$ is precisely the uncalibrated Bayes score f^U. The following lemma establishes an upper bound on the sufficiency gap of the uncalibrated Bayes score in terms of the conditional mutual information between Y and A, conditioning on X. It is a simple consequence of Tao's inequality.

Example B.1 (Sufficiency for uncalibrated Bayes score)**.** *Suppose X and A are discrete $\mathcal{D}$-measurable random variables, and $\mathcal{F}$ is the set of all functions $f : \mathbf{X} \to [0, 1]$. Denote $f^* = \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{L}(f)$. Then, under Assumption 1, $f^* = \mathbb{E}[Y \mid X]$ and*

$$\mathbf{suf}_{f^*}(A) \leq \sqrt{2 \log 2 I(Y; A \mid X)}.$$

Proof. For $f = \mathbb{E}[Y \mid X]$, we have the following identity for $\Delta_f(A)$ by the tower rule:

$$\Delta_f(A) = \mathbb{E}|\mathbb{E}[Y \mid f] - \mathbb{E}[Y \mid f, A]| = \mathbb{E}|\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X, A]|$$

By applying Tao's inequality [Tao, 2006, Ahlswede, 2007], we have that

$$\mathbb{E}|\mathbb{E}[Y \mid X] - \mathbb{E}[Y \mid X, A]| \leq \sqrt{2 \log 2 I(Y; A \mid X)}$$

Note that $I(Y; (X, A) \mid X) = I(Y; A \mid X)$ and the result follows.

□

Lastly, we consider an example when the attribute A is continuous, and there exists a function g which approximately predicts A from X.

Lemma B.2 (Calibration and sufficiency for continuous group attribute A)**.** *Suppose (1) ℓ is the logistic loss, and (2) there exists $g : \mathcal{X} \to \mathcal{A}$ such that $\mathbb{E}[|A - g(X)|] \leq \delta_1$. Let $\tilde{F} = \{f : \mathcal{X} \times \mathcal{A} \to [0, 1] \text{ s.t. } f(x, g(x)) \in \mathcal{F}\}$. Denote $f^* = \arg\inf_{f \in \mathcal{F}} L(f)$ and $\tilde{f} = \arg\inf_{f \in \tilde{F}} L(f)$. Further suppose (3) $\tilde{f}$ is β-Lipschitz in its second argument, that is $\forall x, |\tilde{f}(x, a) - \tilde{f}(x, a')| \leq \beta |a - a'|$ and (4) $\Pr\{\delta_2 < \tilde{f} < \delta_3\} = 1$ for some $\delta_2, \delta_3 \in (0, 1)$. Then,*

$$\max\{\mathbf{suf}_{f^*}(A), \mathbf{cal}_{f^*}(A)\} \leq C \sqrt{\min_{f \in \tilde{\mathcal{F}}} \mathcal{L}(f) - \mathcal{L}(f^B) + \beta \frac{\delta_1}{\min\{\delta_2, 1 - \delta_3\}}},$$

where C is a universal constant.

Proof. By computation, we have

$$\begin{aligned}
\mathcal{L}(f^*) - \mathcal{L}(f^B) &= \mathbb{E}[\ell(f^*(X), Y)] - \mathcal{L}(f^B) \\
&\leq \mathbb{E}[\ell(\tilde{f}(X, g(X)), Y)] - \mathcal{L}(f^B) \\
&\leq \mathbb{E}[\ell(\tilde{f}(X, A), Y)] - \mathcal{L}(f^B) + \beta \frac{\delta_1}{\min\{\delta_2, 1 - \delta_3\}} \\
&= \min_{f \in \tilde{\mathcal{F}}} \mathcal{L}(f) - \mathcal{L}(f^B) + \beta \frac{\delta_1}{\min\{\delta_2, 1 - \delta_3\}}.
\end{aligned}$$

The last inequality follows from the fact that $\forall y,\ \ell(\cdot, y)$ is $\frac{1}{\min\{\delta_2, 1-\delta_3\}}$-Lipschitz on $[\delta_2, \delta_3]$, and that $\tilde{f}$ is only supported on $[\delta_2, \delta_3]$. Then the conclusion follows from Theorem 2.3. □

The above result shows that in the incomplete information setting we may be able to bound the sufficiency gap of the population risk minimizer of class $\mathcal{F}$ by the approximation error for an auxiliary class $\widetilde{\mathcal{F}}$ up to an additional error term that accounts for how well $g(X)$ predicts A. In other words, a score obtained by empirical risk minimization will have low calibration gap with respect to any group attribute A that is sufficiently encoded in the features that are used by the score, X, up to the flexibility allowed by the chosen class $\mathcal{F}$. Thus, our theoretical results also suggest that ERM can achieve simultaneous approximate sufficiency and calibration with respect to all such group attributes.

C Supplementary Material for the Average Sufficiency and Calibration Lower Bound

Before stating the precise version of the lower bound Theorem 2.7, we begin by giving an explicit construction of the hard instance. Let $\mathcal{S}^1 := \{u \in \mathbb{R}^2 : \|u\|_2 = 1\}$ be the circle in $\mathbb{R}^2$. For each $u \in \mathbb{R}^2$, we consider the following affine score functions $f_u : \mathbb{R}^2 \to \mathbb{R}$ and attributes $A_w \in \{-1, 0, 1\}$:

$$f_u(X) := \frac{1}{2} + \frac{\langle u, X\rangle}{4} \quad \text{and} \quad A_w := \operatorname{sign}(\langle X, u\rangle).$$

We note that $f_u(X) \in [\frac{1}{4}, \frac{3}{4}]$ whenever $u, X \in \mathcal{S}^1$, and that our attributes are functions of our features. Lastly, we let $\mathcal{F} := \{f_u : u \in \mathbb{R}^2\}$, and for $\psi \in \mathcal{S}^1$, we let $\mathcal{D}_\theta$ denote the joint distribution where

$$\mathcal{D}_\theta := X \overset{\text{unif}}{\sim} \mathcal{S}^1 \text{ and } Y \mid X = \text{Bernoulli}(f_\theta(X)).$$

Observe that since $A_w = A_w(X)$, we see that the calibrated Bayes score under the distribution $\mathcal{D}_\theta$ is just $f^B(X, A) = \mathbb{E}_{\mathcal{D}_\theta}[Y \mid X, A] = \mathbb{E}_{\mathcal{D}_\theta}[Y \mid X] = f_\theta(X)$, and thus $f^B \in \mathcal{F}$. Lastly, we shall let $\mathbf{suf}_{f;\mathcal{D}_\theta}(A_w)$ denote the sufficiency gap of f with respect to $\mathcal{D}_\theta$, and $\mathbf{cal}_{f;\mathcal{D}_\theta}(A_w)$ the calibration gap of f with respect to $\mathcal{D}_\theta$. We can now state a more precise version of Theorem 1.2:

Theorem C.1 (Precise Lower bound for Sufficiency and Calibration). *Let f_w,$\mathcal{F}$, $\mathcal{D}_\theta$, A_w, and $\mathbf{suf}_{f;\mathcal{D}_\theta}(A_w)$ and $\mathbf{cal}_{f;\mathcal{D}_\theta}$ be as above. Then,*

(a) For any classifier $\widehat{f} \in \mathcal{F}$ trained on a sample $S^n := \{(X_i, Y_i)\}_{i=1}^n$, any $\delta_1 \in (0, 1)$,

$$\mathbb{E}_{\theta, w \overset{\text{unif}}{\sim} \mathcal{S}^1} \Pr_{S^n \sim \mathcal{D}_\theta} \left[\mathbf{suf}_{\widehat{f};\mathcal{D}_\theta}(A_w) \le \frac{1}{4\pi} \min\left\{1, \sqrt{\frac{3\log(1/\delta_1)}{2n}}\right\} \right] \le 1 - \frac{\delta_1}{4}. \tag{22}$$

Moreover, $\mathbf{cal}_{\widehat{f};\mathcal{D}_\theta}(A_w) \ge \mathbf{suf}_{\widehat{f};\mathcal{D}_\theta}(A_w)$ almost surely.

(b) Let $\widehat{f}_n$ denote the ERM under the square loss

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{(X_i, Y_i) \in S^n} (f(X_i) - Y_i)^2.$$

Then (22) holds even when $\mathbb{E}_{\theta, w \overset{\text{unif}}{\sim} \mathcal{S}^1}$ is replaced by a supremum $\sup_{\theta, w \in \mathcal{S}^1}$. Moreover, for any $\delta_2 \in (0, 1)$ and $\theta \in \mathcal{S}^1$,

$$\Pr_{S^n \sim \mathcal{D}_\theta} \left[\mathcal{L}_{\mathcal{D}_\theta}(\widehat{f}_n) - \mathcal{L}^* \le \frac{8 + 6\log(1/\delta_2)}{n} \right] \ge 1 - \delta_2 - 2e^{-\frac{n}{8+4/3}}.$$

where $\mathcal{L}_{\mathcal{D}_\theta}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}_\theta}[(f(X) - Y)^2]$ denotes population risk under the square loss, and where $\mathcal{L}^* = \mathcal{L}_{\mathcal{D}_\theta}(f_\theta)$ denotes the (calibrated) Bayes risk.

The remainder of the section is organized as follows. In Section C.1, we given an overview of the proof strategy for part (a). In Section C.2, we use standard concentration arguments to establish part (b) of the theorem. Sections C.3 and C.2 are devoted to proving the major results whose proofs are omitted in the overview of Section C.1.

C.1 Proof Strategy for Theorem C.1, Part (a)

In this section, we sketch the proof of Theorem C.1, Part (a). Since $\widehat{f} \in \mathcal{F}$, we shall write $\widehat{f} = f_{\widehat{\theta}}$ for some $\widehat{\theta} \in \mathbb{R}^2$. We begin by given a precise characterization of the calibration error:

Lemma C.2. *Let $\theta, w \in \mathcal{S}^1$, and suppose that either $\widehat{\theta} = 0$, or $\mathrm{span}(\widehat{\theta}, w) = \mathbb{R}^2$. Then, for $(X, Y) \sim \mathcal{D}_\theta$,*

$$\mathbf{cal}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) \geq \mathbf{suf}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) = \frac{\sqrt{\Phi(\widehat{\theta}; \theta)}}{2\pi}, \quad \text{where } \Phi(\widehat{\theta}; \theta) = \begin{cases} 1 - \langle \theta, \frac{\widehat{\theta}}{\|\widehat{\theta}\|} \rangle^2 & \widehat{\theta} \neq 0 \\ 1 & \widehat{\theta} = 0 \end{cases}.$$

At the heart of Lemma C.2 is noting that when $\widehat{\theta}$ and w are linearly independent, then $f_{\widehat{\theta}}(X)$ and $A_w(X) = \mathrm{sign}(|\langle X, w \rangle|)$ uniquely determine $X \in \mathcal{S}^1$. Hence, $\mathbb{E}[Y | f_{\widehat{\theta}}(X), A_w(X)] = \mathbb{E}[Y|X] = f_\theta(X)$. In the proof of Lemma C.2, we show that a similar simplification occurs in the case that $\widehat{\theta} = 0$. Because the attribute A_w is independent of the distribution $\mathcal{D}_\theta$, and because $w \overset{\text{unif}}{\sim} \mathcal{S}^1$, we have that, for any θ,

$$\Pr_{w, S^n \sim \mathcal{D}_\theta}[\{\mathrm{span}(w, \widehat{\theta}) = \mathbb{R}^2\} \cup \{\widehat{\theta} = 0\}] = 1, \tag{23}$$

so that the conditions of Lemma C.2 hold with probability one.

Next, we observe that $\Phi(\widehat{\theta}; \theta)$ corresponds to the square norm of the projection of θ onto a direction perpendicular to $\widehat{\theta}$, or equivalently, the square of the sign of the angle between $\widehat{\theta}$ and θ. Note that calibration can occur when the angle between $\widehat{\theta}$ and θ is either close to zero, or close to π-radians; this is in contrast to prediction, where a small loss implies that the angle between $\widehat{\theta}$ and θ is necessarily close to zero. Nevertheless, we can still prove an information theoretic lower bound on the probability that $\Phi(\widehat{\theta}; \theta)$ is small by a reduction to binary hypothesis testing. This is achieved in the next proposition:

Proposition C.3. *For any $n \geq 1$, $\delta \in (0, 1)$, and any estimator $\widehat{\theta}$,*

$$\mathbb{E}_{\theta \overset{\text{unif}}{\sim} \mathcal{S}^1} \Pr_{S^n \sim \mathcal{D}_\theta} \left[\Phi(\widehat{\theta}; \theta) \leq \min \left\{ \frac{1}{2}, \frac{3 \log(1/\delta)}{n} \right\} \right] \leq 1 - \frac{\delta}{4}.$$

The first part of Theorem C.1, Equation (22), now follows immediately from combining the bound in Proposition C.3, (23), and the computation of $\mathbf{suf}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w)$ and $\mathbf{cal}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w)$ in Lemma C.2. The proof of Lemma C.2 and Proposition C.3 are deferred to Sections C.4 and C.3, respectively.

C.2 Proof of Theorem C.1, Part (b): Analysis of $\widehat{f}_n$

We can write $\widehat{f}_n = f_{\widehat{\theta}_{\mathrm{LS}}}$, where

$$\widehat{\theta}_{\mathrm{LS}} := \arg\min_w \sum_{(X_i, Y_i) \in S^n} (f_w(X_i) - Y_i)^2$$

We readily see that the distribution of $\widehat{\theta}_{\mathrm{LS}}$ marginalized over $\theta \overset{\text{unif}}{\sim} \mathcal{S}^1$ is radially symmetric. Hence, the conclusion of (23) holds for any fixed $w \in \mathcal{S}^1$.

Moreover, since $\mathbf{suf}_{f_{\widehat{\theta}_{\mathrm{LS}}}, \mathcal{D}_\theta}(A_w) = \frac{\sqrt{\Phi(\widehat{\theta}_{\mathrm{LS}}; \theta)}}{2\pi}$, and both the least-squares algorithm and the error $\Phi(\cdot, \cdot)$ are radially symmetric, we see that for any t, $\Pr_{S^n \sim \mathcal{D}_\theta}[\mathbf{suf}_{f_{\widehat{\theta}_{\mathrm{LS}}}, \mathcal{D}_\theta}(A_w) \leq t]$ does not depend on $\theta \in \mathcal{S}^1$ either.

It now suffices to prove an upper bound for least squares. We have that

$$\begin{aligned}\mathcal{L}(\widehat{f}_n) - \mathcal{L}^* &= \mathbb{E}\left[\left(f_{\widehat{\theta}_{\mathrm{LS}}}(A)(X) - f_\theta(X)\right)^2\right] \\ &= \mathbb{E}[\langle X, \frac{\widehat{\theta}_{\mathrm{LS}} - \theta}{4}\rangle^2] \\ &= \|\widehat{\theta}_{\mathrm{LS}} - \theta\|^2_{\frac{1}{16}\mathbb{E}[XX]^\top},\end{aligned}$$

where we let $\|x\|_\Sigma^2 := x^\top \Sigma x$. Now, conditioning on $\{X_1, \ldots, X_n\}$, and let $\widehat{\Sigma} := \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$. Observe that $\mathbb{E}[Y \mid X_i] = \langle \theta, \frac{X_i}{4}\rangle$, and $Y_i - \mathbb{E}[Y \mid X_i]$ are independent random variables in $[0, 2]$, so are 1-subgaussian by Hoeffding's inequality. Hence, Hsu et al. [2011, Proposition 1] with $\sigma^2 = 1$ and $d = 2$ implies that

$$\begin{aligned}&\Pr\left[\|\widehat{\theta}_{\mathrm{LS}} - \widehat{\theta}_{\mathrm{LS}}\|^2_{\frac{1}{16}\widehat{\Sigma}} \le \frac{4 + 3\log(1/\delta)}{n} \mid \{X_1, \ldots, X_n\}\right] \\ &\overset{(i)}{\ge} \Pr\left[\|\widehat{\theta}_{\mathrm{LS}} - \widehat{\theta}_{\mathrm{LS}}\|^2_{\frac{1}{16}\widehat{\Sigma}} \le \frac{2 + 2\sqrt{2\log(1/\delta)} + 2\log(1/\delta)}{n} \mid \{X_1, \ldots, X_n\}\right] \ge 1 - \delta.\end{aligned}$$

where (i) uses the elementary inequality $ab \le \frac{a^2+b^2}{2}$. Lastly, we note that $\mathbb{E}[XX^\top] = \frac{1}{2}I$, so on the event $\lambda_{\min}(\widehat{\Sigma}) \ge \frac{1}{4}$, we have $\left\|\widehat{\theta}_{\mathrm{LS}} - \theta\right\|^2_{\frac{1}{16}\mathbb{E}[XX^\top]} \le \frac{1}{2}\|\widehat{\theta}_{\mathrm{LS}} - \theta\|^2_{\frac{1}{16}\widehat{\Sigma}}$. To this end, define $M_i = \mathbb{E}[XX^\top] - X_i X_i^\top = \frac{1}{2}I - X_i X_i^\top$. Note that $\lambda_{\max}(M_i) \le \frac{1}{2}$ and $\mathbb{E}[M_i^2] = \frac{1}{4}I$. Hence, by the Matrix Bernstein inequality Tropp [2015, Theorem 6.6.1], we have

$$\Pr\left[\lambda_{\min}(\widehat{\Sigma}) \le \frac{1}{4}\right] = \Pr\left[\lambda_{\max}(\sum_{i=1}^n M_i) \ge \frac{n}{4}\right] \le 2e^{-\frac{t^2/2}{(n/4)+(t/6)}}\big|_{t=\frac{n}{4}} = 2e^{-\frac{n}{8+4/3}}.$$

Putting pieces together, we conclude that

$$\Pr\left[\mathbb{E}\left[\left(f_{\widehat{\theta}_{\mathrm{LS}}}(A)(X) - f_\theta(X)\right)^2\right] \le \frac{8 + 6\log(1/\delta)}{n}\right] \ge 1 - \delta - 2e^{-\frac{n}{8+4/3}}.$$

C.3 Proof of Information Theoretic Bound, Proposition C.3

Let $R_\psi : \mathbb{R}^2 \to \mathbb{R}^2$ denote the linear operator corresponding to rotation by an angle $\psi \in [0, 2\pi]$. Our strategy will be to show that for any $w \in \mathcal{S}^1$, for

$$\epsilon(n) = \sqrt{\min\left\{\frac{1}{2}, \frac{3\log(1/\delta)}{n}\right\}}, \tag{24}$$

and some angle $\psi = \psi(n)$ depending on n, we have

$$\frac{1}{2}\left(\Pr[\Phi(\widehat{\theta}; \psi) \le \epsilon(n)^2] + \Pr[\Phi(\widehat{\theta}; R_\psi \psi)] \le \epsilon(n)^2\right) \le 1 - \frac{\delta}{4} \tag{25}$$

Indeed, if (25) holds, then we may express can express $\psi = R_\phi e_1$ where $e_1 = (1, 0)$ for some $\phi \in [0, 2\pi]$. Thus, taking an expectation over $\phi \overset{\mathrm{unif}}{\sim} [0, 2\pi]$, we observe that

$$\mathbb{E}_{\phi \overset{\mathrm{unif}}{\sim}[0,2\pi]} \Pr\left[\Phi(\widehat{\theta}; R_\phi e_1) \le \epsilon(n)^2\right] = \mathbb{E}_{\theta \overset{\mathrm{unif}}{\sim} \mathcal{S}^1} \Pr\left[\Phi(\widehat{\theta}; \theta) \le \epsilon(n)^2\right],$$

and similarly

$$\begin{aligned}\mathbb{E}_{\phi \overset{\text{unif}}{\sim} [0,2\pi]} \Pr\left[\Phi(\widehat{\theta}; R_\psi R_\phi e_1) \le \epsilon(n)\right] &= \mathbb{E}_{\phi \overset{\text{unif}}{\sim} [0,2\pi]} \Pr\left[\Phi(\widehat{\theta}; R_{\phi+\psi} e_1) \le \epsilon(n)\right] \\ &= \mathbb{E}_{\widetilde{\phi} \overset{\text{unif}}{\sim} [0,2\pi]} \Pr\left[\Phi(\widehat{\theta}; R_{\widetilde{\phi}} w) \le \epsilon(n)^2\right].\end{aligned}$$

Hence,

$$\begin{aligned}1 - \frac{\delta}{4} &\ge \frac{1}{2}\mathbb{E}_{\phi \overset{\text{unif}}{\sim} [0,2\pi]} \left[\Pr\left[\Phi(\widehat{\theta}; R_\phi w) \le \epsilon(n)^2\right] + \Pr\left[\Phi(\widehat{\theta}; R_\psi R_\phi e_1)\right] \le \epsilon(n)^2\right] \\ &= \frac{1}{2} \cdot 2\mathbb{E}_{\theta \overset{\text{unif}}{\sim} \mathcal{S}_1} \Pr\left[\Phi(\widehat{\theta}; \theta) \le \epsilon(n)^2\right], \quad \text{as needed.}\end{aligned}$$

We now turn to proving (25). By rotation invariance argument, it suffices to prove the inequality for $w = e_1 = (1, 0)$. We now fix an $\epsilon = \epsilon(n)$ as in (24) to be chosen later, and choose $\psi = \arccos(1 - 2\epsilon^2)$. Note that $\epsilon \in (0, \frac{1}{2}]$ implies $\psi \in (0, \pi/2]$.

We construct two alternative instances $\theta^{(1)} = e_1$, and let $\theta^{(2)} = e_1 \cos\psi + e_2 \sin\psi$. We will establish a lower bound on the problem of testing between $\theta = \theta^{(1)}$ and $\theta = \theta^{(2)}$, and then translate this into a bound on $\Phi(\widehat{\theta}; \cdot)$. The first step is a KL-divergence computation established in Section C.3.1:

Lemma C.4. *There exists a constant $K > 0$ such that, if $(\mathcal{D}_\theta)^{\otimes n}$ denote the distribution of n i.i.d. samples from $\mathcal{D}_\theta$, then*

$$\mathrm{KL}((\mathcal{D}_{\theta_1})^{\otimes n}, (\mathcal{D}_{\theta_2})^{\otimes n}) \le \frac{n}{12}\|\theta_1 - \theta_2\|_2^2.$$

In our setting, we see that

$$\|\theta_1 - \theta_2\|_2^2 = (1 - \cos\psi)^2 + \sin^2\psi = 1 + \cos^2\psi + \mathrm{sign}^2\psi - 2\cos\psi = 2(1 - \cos\psi) = 4\epsilon^2.$$

Hence, we have that $\mathrm{KL}((\mathcal{D}_{\theta_1})^{\otimes n}, (\mathcal{D}_{\theta_2})^{\otimes n}) \le n\frac{\epsilon^2}{3}$. Therefore, given any estimator $\widehat{i}$ of i, the proof of [Theorem 2.2.iii in Tsybakov [2008]] reveals that

$$\frac{1}{2}\sum_{i \in \{1,2\}} \Pr_{(\mathcal{D}_{\theta_i})^{\otimes n}}\left[\{\widehat{i} \ne i\}\right] \ge \frac{1}{4}e^{-\frac{n\epsilon^2}{3}}.$$

In particular, since $\epsilon = \epsilon(n)^2 \le \frac{3\log(1/\delta)}{n}$, as in (24), and considering the complement of $\{\widehat{i} \ne i\}$, we have

$$\frac{1}{2}\sum_{i \in \{1,2\}} \Pr_{(\mathcal{D}_{\theta_i})^{\otimes n}}\left[\{\widehat{i} = i\}\right] \le 1 - \frac{1}{4}e^{-\log(1/\delta)} = 1 - \frac{\delta}{4} \tag{26}$$

Lastly, we show how a small value of $\Phi(\widehat{\theta}; \theta_i)$ yields an accurate estimator of $\widehat{i}$. Given an estimator $\widehat{\theta}$, we define the estimator of θ_i give $\theta_{\widehat{i}}$, where $\widehat{i}$ is given by:

$$\widehat{i} \in \arg\min_{i \in \{1,2\}} \Phi(\theta_i; \widehat{\theta}),$$

where we arbitrarily choose $\widehat{i} = 1$ if both values of i attain the same value in the display above. The following lemma, proved in Section C.3.2, shows that gives a reduction from estimating i to obtaining a small value of $\Phi(\theta_i, \widehat{\theta})$:

Lemma C.5. *For $\psi \in [0, \frac{\pi}{2}]$, $\Phi(\theta_i, \widehat{\theta}) < \sin^2 \frac{\psi}{2}$ implies that $\widehat{i} = i$.*

Hence, combining Lemma C.5 with (26), we have

$$\frac{1}{2} \sum_{i \in \{1,2\}} \Pr[\Phi(\theta_i, \widehat{\theta}) < \sin \frac{\psi}{2}] \leq 1 - \frac{\delta}{4}$$

Lastly, we find that as $\psi \in (0, \frac{\pi}{2})$, $\sin \frac{\psi}{2} = \sqrt{\frac{1-\cos\psi}{2}} = \sqrt{\frac{2\epsilon^2}{2}} = \epsilon$, thereby concluding the proof.

C.3.1 Proof of Lemma C.4

By the tensorization of the KL-divergence,

$$\begin{aligned} \mathrm{KL}\left((\mathcal{D}_{\theta_1})^{\otimes n}, (\mathcal{D}_{\theta_2})^{\otimes n}\right) &= n\mathrm{KL}\left(\mathcal{D}_{\theta_1}, \mathcal{D}_{\theta_2}\right) \\ &= n\mathbb{E}_{X \overset{\text{unif}}{\sim} \mathcal{S}^1} \mathrm{KL}\left(\text{Bernoulli}(f_{\theta_1}(X)), \text{Bernoulli}(f_{\theta_2}(X))\right), \end{aligned}$$

Now we use a standard Taylor-expansion upper bound on the KL-divergence between two Bernoulli random variables (see, e.g. Lemma E.1 in Simchowitz et al. [2016]):

Lemma C.6. *Let $p, q \in (0, 1)$. Then,*

$$\mathrm{KL}\left(\text{Bernoulli}(p), \text{Bernoulli}(q)\right) \leq \frac{(p-q)^2}{2\min\{p(1-p), q(1-q)\}}.$$

In our setting,

$$f_{\theta_i}(X) = \frac{1}{2} + \frac{\langle \theta_i, X \rangle}{4} \in \left[\frac{1}{4}, \frac{3}{4}\right] \quad \text{because } |\langle X, \theta_i \rangle| \leq \frac{\|\theta_i\|\|x\|}{4} = \frac{1}{4}.$$

Hence, since $2\min_{p \in [1/4, 3/4]} p(1-p) = 2 \cdot \frac{1}{4} \cdot \frac{3}{4} = 3/8$,

$$\mathrm{KL}\left(\text{Bernoulli}(f_{\theta_1}(X)), \text{Bernoulli}(f_{\theta_2}(X))\right) \leq \frac{8\left\|f_w(X) - f_{w'}(X)\right)\|_2^2}{3}$$

Hence,

$$\begin{aligned} KL\left((\mathcal{D}_{\theta_1})^{\otimes n}, (\mathcal{D}_{\theta_2})^{\otimes n}\right) &\leq n \cdot \frac{8}{3} \mathbb{E}_{X \overset{\text{unif}}{\sim} \mathcal{S}^1} \left\|f_{\theta_1}(X) - f_{\theta_2}(X))\right\|_2^2 \\ &= \frac{8n}{3} \mathbb{E}_{X \overset{\text{unif}}{\sim} \mathcal{S}^1} \left\| \left\langle \frac{\theta_1 - \theta_2}{4}, X \right\rangle \right\|_2^2 \\ &= \frac{n}{6} (\theta_1 - \theta_2)^\top \mathbb{E}_{X \sim \mathcal{S}^1}[XX^\top](\theta_1 - \theta_2) = \frac{n}{12} \|\theta_1 - \theta_2\|_2^2. \end{aligned}$$

C.3.2 Proof of Lemma C.5

By assumption, we have that $\Phi(\theta_i, \widehat{\theta}) < \sin^2 \frac{\psi}{2}$ for some $\psi \in [0, \frac{\pi}{2}]$. Since $\psi \leq \frac{\pi}{2}$, $\sin^2 \frac{\psi}{2} < 1$, so $\Phi(\theta_i; \widehat{\theta}) < 1$, ruling out the case $\widehat{\theta} = 0$. Thus, we may write can write

$$\frac{\widehat{\theta}}{\|\widehat{\theta}\|} = e_1 \cos\phi + e_2 \sin\phi$$

for some $\phi \in [-\pi, \pi]$. Since $\widehat{\theta} \neq 0$, $\Phi(\theta_i; \widehat{\theta})$ corresponds to the square norm of the projection of θ_i onto a direction perpendicular to $\widehat{\theta}$. Therefore,

$$\Phi(\theta_1; \widehat{\theta}) = \sin^2 \phi \text{ and } \Phi(\theta_2; \widehat{\theta}) = \sin^2(\phi - \psi).$$

We shall show that if $\Phi(\theta_1; \widehat{\theta}) < \sin^2 \frac{\psi}{2}$, then $\widehat{i} = 1$; proving that $\Phi(\theta_2; \widehat{\theta}) < \sin^2 \frac{\psi}{2}$ implies $\widehat{i} = 2$ is analogous. To this end, suppose that $\sin^2 \phi = \Phi(\theta_1; \widehat{\theta}) < \sin^2 \frac{\psi}{2}$. We consider three cases, and show that in each case, $\sin^2(\phi - \psi) \geq \sin^2 \frac{\psi}{2}$.

1. Case (a): $|\phi| < \frac{\psi}{2}$. Then, we have that $\psi - \phi \in (\frac{\psi}{2}, \frac{3\psi}{2})$. Since $\psi \leq \frac{\pi}{2}$, $\min_{\varphi \in [\frac{\psi}{2}, \frac{3\psi}{2}]} \sin^2 \varphi = \sin^2 \frac{\psi}{2}$, whence $\sin^2 \frac{\psi}{2} < \sin^2(\phi - \psi)$.

2. Case (b.1): $\phi \in (\pi - \frac{\psi}{2}, \pi]$. Then, we have that $\phi - \psi \in (\pi - \frac{3}{2}\psi, \pi - \psi]$. Now, if $\phi - \psi \in [\frac{\pi}{2}, \pi - \psi)$, we have that $\sin^2(\phi - \psi) \in [\sin^2 \psi, 1]$, so that $\sin^2(\phi - \psi) \geq \sin^2 \frac{\psi}{2}$. On the other hand, if $\phi - \psi \in [\pi - \frac{3\psi}{2}, \frac{\pi}{2}]$, then since $\psi \leq \frac{\pi}{2}$, we have $\sin^2(\phi - \psi] \in [\frac{\pi}{4}, \pi/2]$. Thus, $\sin^2(\phi - \psi) \geq \sin^2 \frac{\pi}{4} \geq \sin^2 \frac{\psi}{2}$.

3. Case (b.2): $\phi \in [-\pi, -\pi + \frac{\psi}{2})$. Then, we have that $\phi - \psi \in [-\pi - \psi, \pi - \frac{\psi}{2})$, and the rest is similar to (b.1).

C.4 Proof of Calibration Error Computation, Lemma C.2

Recall that our goal is to establish that

$$\mathbf{cal}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) \geq \mathbf{suf}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) = \frac{\sqrt{\Phi(\widehat{\theta}; \theta)}}{2\pi}, \quad \text{where } \Phi(\widehat{\theta}; \theta) = \begin{cases} 1 - \langle \theta, \frac{\widehat{\theta}}{\|\widehat{\theta}\|} \rangle^2 & \widehat{\theta} \neq 0 \\ 1 & \widehat{\theta} = 0 \end{cases}.$$

We begin with a more involved calculation to compute $\mathbf{suf}_f(A)$; we shall lower bound $\mathbf{cal}_f(A)$ at the end of the section.

Bound for $\mathbf{suf}_f(A)$: We shall show that if either $\text{span}(\{w, \widehat{\theta}\}) = \mathbb{R}^2$ or $\widehat{\theta} = 0$, then there is a unit vector $v \in \mathcal{S}^1$ for which

$$\mathbf{suf}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) = \frac{\sqrt{\Phi(\widehat{\theta}; \theta)}}{4} \cdot \mathbb{E}_{X \overset{\text{unif}}{\sim} \mathcal{S}^1}[|\langle v, X \rangle|].$$

This is enough to conclude the proof, since

$$\mathbb{E}_{X \overset{\text{unif}}{\sim} \mathcal{S}^1}[|\langle v, X \rangle|] = \frac{1}{2\pi} \int_0^{2\pi} |\sin \psi| d\psi \;=\; \frac{1}{\pi} \int_0^{\pi} \sin \psi d\psi = \frac{2}{\pi}.$$

First, suppose that $\widehat{\theta} \neq 0$. Choose an orthonormal basis $\{e_1, e_2\}$ so that $\widehat{\theta} = \|\widehat{\theta}\| e_1$. Then, we can write

$$X = X_1 e_1 + X_2 e_2 \ ,$$

where $X_i = \langle X_i, e_i \rangle$. Then, letting $\theta = \theta[1] e_1 + \theta[2] e_2$, we see that

$$\theta[2] = \sqrt{\Phi(\widehat{\theta}; \theta)},$$

and we have

$$f_\theta(X) = \langle \theta, X\rangle + \frac{1}{2} = \frac{1}{4}X_1\theta[1] + \frac{1}{4}X_2\theta[2] + \frac{1}{2}.$$

First, suppose that $\widehat{\theta} \neq 0$. Then, since $f_{\widehat{\theta}}(X) = \frac{1}{2} + \|\widehat{\theta}\| \cdot X_1$ is in bijection with X_1, and since

$$\mathbb{E}[X_2|X_1] = 0 \text{ for } (X_1, X_2) \overset{\text{unif}}{\sim} \mathcal{S}^1,$$

we have

$$\mathbb{E}[f_\theta(X) \mid f_{\widehat{\theta}}(X)] = \mathbb{E}[f_\theta(X) \mid X_1] \;=\; \frac{1}{2} + \frac{1}{4}X_1\theta[1] + \frac{1}{4}\theta[2]\mathbb{E}[X_2 \mid X_1] \;= \frac{1}{2} + \frac{1}{4}X_1\theta[1].$$

Moreover, if w and $w = \|w\|e_1$ are linearly independent, then since $X \in \mathcal{S}_1$, $A_w = \text{sign}(\langle w, X\rangle)$ and $X_1 = \langle e_1, X\rangle$ uniquely determine X. Hence,

$$\mathbb{E}[f_\theta(X) \mid f_{\widehat{\theta}}, A_w] = \mathbb{E}[f_\theta(X) \mid X] \;=\; f_\theta(X) = \frac{1}{2} + \frac{1}{4}X_1\theta[1] + \frac{1}{4}\theta[2]X_2.$$

Thus,

$$\mathbb{E}[f_\theta(X) \mid f_w(X), A] - \mathbb{E}[f_\theta(X) \mid f_w(X)] = \frac{\theta[2]X_2}{4}$$

Hence, we conclude that

$$\textbf{suf}_f(A) = \frac{|\theta[2]|}{4} \cdot \mathbb{E}[|X_2|] = \frac{\sqrt{\Phi(\widehat{\theta};\theta)}}{4}\mathbb{E}[|\langle e_2, X\rangle|].$$

We now address the edge-case $\widehat{\theta} = 0$. Since $f_{\widehat{\theta}}(X) = 0$ for all X, $\mathbb{E}[Y \mid f_{\widehat{\theta}}] = \mathbb{E}[Y] = \frac{1}{2}$, and $\mathbb{E}[Y \mid f_{\widehat{\theta}}, A_w] = \mathbb{E}[Y \mid A_w]$. To compute $\mathbb{E}[Y \mid A_w]$, let $e_1 = w$, and let e_2 be such that $\{e_1, e_2\}$ form an orthonormal basis, and write $X = X_1e_1 + X_2e_2$. Then,

$$\begin{aligned}
\mathbb{E}[X \mid A] &= \mathbb{E}[X_1 \mid \text{sign}(X_1)] + \mathbb{E}[X_2 \mid \text{sign}(X_1)] \\
&= \mathbb{E}[X_1 \mid \text{sign}(X_1)] && \text{(since } X_2 \perp \text{sign}(X_1)\text{)} \\
&= \text{sign}(X_1)\mathbb{E}[\text{sign}(X_1)X_1 \mid \text{sign}(X_1)] \\
&= \text{sign}(X_1)\mathbb{E}[|X_1| \mid \text{sign}(X_1)] \;=\; \text{sign}(X_1)\mathbb{E}[|X_1|].
\end{aligned}$$

Hence, $\mathbb{E}[Y \mid A] = \frac{1}{2} + \frac{1}{4}\langle \theta, \mathbb{E}[X \mid A]\rangle = \frac{\text{sign}(X_1)\mathbb{E}[|X_1|]}{4}$. Therefore,

$$\begin{aligned}
\textbf{suf}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) &= \mathbb{E}\left|\mathbb{E}[f_\theta(X) \mid f_{\widehat{\theta}}(X), A_w] - \mathbb{E}[f_\theta(X) \mid f_{\widehat{\theta}}(X)]\right| \\
&= \mathbb{E}\left|\frac{1}{2} + \frac{\text{sign}(X_1)\mathbb{E}[|X_1|]}{4} - \frac{1}{2}\right| = \frac{1}{4}\mathbb{E}[|X_1|] = \frac{\sqrt{\Phi(\widehat{\theta};\theta)}}{4}\mathbb{E}[|\langle w, X\rangle|], \qquad (27)
\end{aligned}$$

where we recall the convention $\Phi(\widehat{\theta};\theta) = 1$ if $\widehat{\theta} = 0$.

Bound for cal: First consider a function $f = f_{\widehat{\theta}}$ for some $\widehat{\theta} \in \mathbb{R}^2$. Suppose first that $\widehat{\theta} \neq 0$. As established above, $\mathbb{E}[Y|A_w, f_{\widehat{\theta}}] = f_\theta$ almost surely, so we have we have

$$\begin{aligned}
\textbf{cal}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) &= \mathbb{E}_{X\sim\mathcal{S}^1}|f_{\widehat{\theta}}(X) - \mathbb{E}[Y|f_{\widehat{\theta}}(X), A_w]| \\
&= \mathbb{E}_{X\sim\mathcal{S}^1}|f_{\widehat{\theta}}(X) - f_\theta(X)| \\
&= \mathbb{E}_{X\sim\mathcal{S}^1}|\frac{1}{4}\langle \widehat{\theta} - \theta, X\rangle| \\
&\overset{(i)}{=} \frac{\|\widehat{\theta} - \theta\|_2}{4}\mathbb{E}_{X\sim\mathcal{S}^1}|\langle e_1, X\rangle| \\
&\overset{(ii)}{=} \frac{1}{2\pi}\|\widehat{\theta} - \theta\|_2,
\end{aligned}$$

where (i) uses the fact that X has a rotation invariant distribution, and (ii) uses the computation $\mathbb{E}_{X\sim\mathcal{S}^1}|\langle e_1, X\rangle| = 2/\pi$ performed above. To let (e_1, e_2) be an orthonormal basis for $\mathbb{R}^2$ for which $\widehat{\theta} = \|\widehat{\theta}\|e_1$, and let $w_* = w_{1,*}e_1 + w_{2,*}e_2$ as above. Then

$$\|\widehat{\theta} - \theta\|_2 = \sqrt{(\|w\|_2 - w_{1,*})^2 + w_{2,*}^2} \geq w_{2,*} = \sqrt{\Phi(\widehat{\theta}, \theta)},$$

as needed.

If $\widehat{\theta} = 0$, then as show above $\mathbb{E}[Y \mid f_{\widehat{\theta}}, A_w] = \mathbb{E}[Y \mid A_w] = \frac{\mathrm{sign}(X_1)\mathbb{E}[|X_1|]}{4}$, and $f_{\widehat{\theta}}(X) = \frac{1}{2}$. Hence,

$$\begin{aligned}
\mathbf{cal}_{f_{\widehat{\theta}}, \mathcal{D}_\theta}(A_w) &= \mathbb{E}_{X\sim\mathcal{S}^1}|f_{\widehat{\theta}}(X) - \mathbb{E}[Y|f_{\widehat{\theta}}(X), A_w]| \\
&= \mathbb{E}_{X\sim\mathcal{S}^1}|\frac{1}{2} - \frac{\mathrm{sign}(X_1)\mathbb{E}[|X_1|]}{4}| \\
&= \frac{\sqrt{\Phi(\widehat{\theta}; \theta)}}{4}\mathbb{E}[|\langle w, X\rangle|] \quad \text{(by (27))},
\end{aligned}$$

which is equal to $\frac{\sqrt{\Phi(\widehat{\theta};\theta)}}{2\pi}$ by the computation of $\mathbb{E}[|\langle w, X\rangle|]$ above.

D Supplementary Material for the Per-group Sufficiency and Calibration Lower Bound

This section contains a formal analogue of Theorem 2.8. Again, we begin with a construction of the joint distribution over features, attributes and labels before stating the precise result.

D.1 Construction:

We construct a "product" of two independent instances of Theorem C.1. We will put "bars" over quantities related to the product distribution, function class, etc.., and use α and β to denote each component of the product.

Let $\mathcal{D}_\cdot$ denote the distribution from the construction in Theorem C.1. We will draw $\theta^\alpha, \theta^\beta$ independently from $\mathcal{S}^1$. We let $\overline{\theta} = (\theta^\alpha, \theta^\beta)$, and define $(\overline{X}, Y, Z) \sim \overline{\mathcal{D}}_{\overline{\theta}}$ as follows:

1. Let $Z \sim \text{Bernoulli}(p)$.
2. If $Z = 1$, draw $(X^\alpha, Y) \sim \mathcal{D}_{\theta^\alpha}$, otherwise draw $(X^\beta, Y) \sim \mathcal{D}_{\theta^\beta}$.
3. Let $\overline{X} = (X^\alpha, 0) \in \mathbb{R}^4$ if $Z = 1$; otherwise set $\overline{X} = (0, X^\beta) \in \mathbb{R}^4$.

Note that Z can be determined from $\overline{X}$ by looking at which coordinate is 0. Further we define:

1. $\overline{f}_{\widehat{\overline{\theta}}}(\overline{X}) = \frac{1}{2} + \frac{1}{4}\langle\widehat{\overline{\theta}}, \overline{X}\rangle$ for $\overline{X}, \widehat{\overline{\theta}} \in \mathbb{R}^4$, and $\overline{\mathcal{F}} := \{\overline{f}_{\widehat{\overline{\theta}}} : \widehat{\overline{\theta}} \in \mathbb{R}^4\}$.
2. The loss function $\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}) = \mathbb{E}_{(\overline{X},Y)\sim\overline{\mathcal{D}}_{\overline{\theta}})}((\overline{f}(X) - \mathbb{E}[Y|\overline{X}])^2 = \mathbb{E}_{(\overline{X},Y)\sim\overline{\mathcal{D}}_{\overline{\theta}})}((\overline{f}(X) - \overline{f}_{\overline{\theta}})^2$.
3. We let $\mathcal{L}_*$ denote the Bayes loss $\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\overline{\theta}})$, which we note does not depend on $\overline{\theta}$.

Lastly for $\overline{w} = (w^\alpha, w^\beta) \in \mathcal{S}^1 \times \mathcal{S}^1$, we define our discrete attribute $\overline{\mathcal{A}}_{\overline{w}}(\overline{X})$ which takes 4 values.

$$\overline{\mathcal{A}}_{\overline{w}}(\overline{X}) = (Z, \text{sign}(\langle \overline{w}, \overline{X} \rangle)) \in \{0,1\} \times \{-1,1\}.$$

Here, we we note that with probability 1, $\text{sign}(\langle \overline{w}, \overline{X} \rangle) = Z\text{sign}(\langle w^\alpha, \overline{X}^\alpha \rangle) + (1-Z)\text{sign}(\langle w^\beta, \overline{X}^\beta \rangle) \neq 0$. In the statement of the theorem, we map $\overline{\mathcal{A}}_{\overline{w}}(\overline{X}) \to \{1, \ldots, 4\}$ via the bijection $(Z, \text{sign}(\langle \overline{w}, \overline{X} \rangle)) \mapsto \frac{1+\text{sign}(\langle \overline{w}, \overline{X} \rangle))}{2} + 2(1-Z)$, which maps the $Z = 1$ attributes to $\{1, 2\}$.

For now, using the $\{0,1\} \times \{-1,1\}$ attributes will be more transparent. Lastly, we see that for attributes $a \in \{(0,-1),(0,1)\}$, and any classifier $\overline{f}_{\widehat{\theta}}$ that

$$\begin{aligned}
&\frac{1}{2}\left(\mathbf{suf}_{\overline{f}_{\widehat{\theta}};\overline{\mathcal{D}}_{\overline{\theta}}}((0,-1);\overline{\mathcal{A}}_{\overline{w}}) + \mathbf{suf}_{\overline{f}_{\widehat{\theta}};\overline{\mathcal{D}}_{\overline{\theta}}}(\mathcal{A}_{\overline{w}}((0,-1);\overline{X})\right) \\
&= \mathbb{E}[\left|\mathbb{E}[Y|\overline{f}_{\widehat{\theta}}] - \mathbb{E}[Y|\overline{f}_{\widehat{\theta}}, \text{sign}(\langle \overline{w}, \overline{X} \rangle)\right| \mid Z = 1] \\
&= \mathbb{E}[\left|\mathbb{E}[Y|f_{\widehat{\theta}^\alpha}] - \mathbb{E}[Y|f_{\widehat{\theta}^\alpha}, \text{sign}(\langle w^\alpha, X^\alpha \rangle)\right| \mid Z = 1] \\
&= \mathbf{suf}_{f_{\widehat{\theta}^\alpha};\mathcal{D}_{\theta^\alpha}}(A_{w^\alpha}),
\end{aligned}$$

that is, the calibration term from Theorem C.1. Hence, by Lemma C.2 in the proof of Theorem C.1, we find that

$$\max_{a \in \{(0,-1),(0,1)\}} \mathbf{suf}_{\overline{f}_{\widehat{\theta}};\overline{\mathcal{D}}_{\overline{\theta}}}(a;\overline{\mathcal{A}}_{\overline{w}}) \geq \mathbf{suf}_{f_{\widehat{\theta}^\alpha};\mathcal{D}_{\theta^\alpha}}(A_{w^\alpha}) = \frac{\sqrt{\Phi(\widehat{\theta}^\alpha;\theta^\alpha)}}{2\pi}. \tag{28}$$

A similar argument shows that

$$\max_{a \in \{(0,-1),(0,1)\}} \mathbf{cal}_{\overline{f}_{\widehat{\theta}};\overline{\mathcal{D}}_{\overline{\theta}}}(a;\overline{\mathcal{A}}_{\overline{w}}) \geq \mathbf{cal}_{f_{\widehat{\theta}^\alpha};\mathcal{D}_{\theta^\alpha}}(A_{w^\alpha}) = \frac{\sqrt{\Phi(\widehat{\theta}^\alpha;\theta^\alpha)}}{2\pi}. \tag{29}$$

D.2 Statement of the Exact Theorem

We now state the technical version of Theorem D.1; we remark that the p in the theorem as stated previously corresponds to $p/2$ in the following statement:

Theorem D.1. *Fix any $p \in (0, 1/2)$, and let $\overline{\mathcal{F}}$, $A_{\overline{w}}$, $\overline{\mathcal{D}}_{\overline{\theta}}$ be as above, indexed by $\overline{w}, \overline{\theta} \in \mathcal{S}^1 \times \mathcal{S}^1$. Further, write $\overline{w}, \overline{\theta} \sim \mathcal{S}^1 \times \mathcal{S}^1$ to denote that $\overline{w}$ and $\overline{\theta}$ are drawn independently from the uniform distribution on $\mathcal{S}^1 \times \mathcal{S}^1$. Then, for any $\overline{w} \in \mathcal{S}^1 \times \mathcal{S}^1$, $\Pr[A_{\overline{w}} = 1] = \Pr[A_{\overline{w}} = 2] = p/2$ and*

(a) For any classifier $\widehat{f} \in \overline{\mathcal{F}}$ trained on a sample $S^n := \{(X_i, Y_i)\}_{i=1}^n$, any $\delta \in (0,1)$,

$$\begin{aligned}
\mathbb{E}_{\overline{\theta},\overline{w} \sim \mathcal{S}^1 \times \mathcal{S}^1} \Pr_{S^n \sim \overline{\mathcal{D}}_{\overline{\theta}}} \left[\max_{a \in \{1,2\}} \min\{\mathbf{cal}_{\widehat{f};\overline{\mathcal{D}}_{\overline{\theta}}}(a;A_{\overline{w}}), \mathbf{suf}_{\widehat{f};\overline{\mathcal{D}}_{\overline{\theta}}}(a;W_{\overline{w}})\} \leq c_1 \min\left\{1, \sqrt{\frac{\log(1/\delta_1)}{pn}}\right\}\right] \\
\leq 1 - c_0\delta. \quad (30)
\end{aligned}$$

(b) Let $\widehat{f}_n$ denote the ERM under the square loss

$$\widehat{f}_n := \arg\min_{f \in \mathcal{F}} \sum_{(X_i,Y_i) \in S^n} (f(X_i) - Y_i)^2.$$

Then (30) holds even when $\mathbb{E}_{\overline{\theta},\overline{w}\overset{\text{unif}}{\sim}\mathcal{S}^1\times\mathcal{S}^1}$ *is replaced by a supremum* $\sup_{\overline{\theta},\overline{w}\in\mathcal{S}^1\times\mathcal{S}^1}$*. Moreover, for any* $\delta_2 \in (0, 1/4)$ *and* $\overline{\theta} \in \mathcal{S}^1 \times \mathcal{S}^1$*,*

$$\Pr_{S^n\sim\overline{\mathcal{D}}_{\overline{\theta}}}\left[\mathcal{L}_{\overline{\mathcal{D}}_{\overline{\theta}}}(\widehat{f}_n) - \mathcal{L}^* \le c_2\frac{\log(1/\delta_2)}{n}\right] \ge 1 - \delta_2 - 4e^{-c_3pn},$$

where $\mathcal{L}_{\overline{\mathcal{D}}_{\overline{\theta}}}(f) = \mathbb{E}_{(X,Y)\sim\overline{\mathcal{D}}_{\overline{\theta}}}[(f(X) - Y)^2]$ *denotes population risk under the square loss, and where* $\mathcal{L}^* = \mathcal{L}_{\overline{\mathcal{D}}_{\overline{\theta}}}(\overline{f}_{\overline{\theta}})$ *denotes the (calibrated) Bayes risk.*

D.3 Proof of Theorem D.1

Learning Setup: Let $\mathcal{S} = \{(\overline{X}_i, Y_i, Z_i)\}_{i=1}^n$ be a sample drawn *i.i.d* from $\mathcal{D}_\theta$, and define the subsamples $\mathcal{S}_\alpha := \{(\overline{X}_i^\alpha, Y_i) : Z_i = 1\}$ and $\mathcal{S}_\beta = \{(\overline{X}_i^\beta, Y_i) : Z_i = 0\}$, and sample numbers $n_\alpha := |\mathcal{S}_\alpha|$ and $n_\beta := |\mathcal{S}_b|$. Then conditional on n_α (or equivalently, on n_β), $\mathcal{S}_a$ has the distribution of a sample of size n_α from $\mathcal{D}_{\theta^\alpha}$, where $\mathcal{D}_w$ is the distribution from the 2-group lower bound, Theorem C.1, and is independent of $\mathcal{S}_\beta$. Lastly, we define the "Chernoff' 'event

$$\mathcal{E}_{\mathsf{Cher}} := \{n_\alpha \ge \frac{1}{2}pn \text{ and } n_\beta \ge \frac{1}{2}(1-p)n\}.$$

And we note that $\Pr[\mathcal{E}_{\mathsf{Cher}}] \ge 1 - 2e^{-\Omega(\min\{pn,(1-pn)\})} \ge 1 - 2e^{-c_1pn}$ for some constant c_1 by combining Chernoff bounds for n_α and n_β. We now can prove part 1 and part 2 of the proposition separately.

Part 1: Information Theoretic Lower Bound: Let's condition on n_α. Note then that $\mathcal{S}_\beta$ contains no information about θ^α, since the prior on θ^α and θ^β are independent. Thus, the information theoretic lower bound, Proposition C.3, implies we have that for the α-component of our estimator, $\widehat{\overline{\theta}}^\alpha$, must satisfy the bound:

$$\mathbb{E}_{\theta^\alpha\overset{\text{unif}}{\sim}\mathcal{S}^1}\Pr_{\mathcal{S}_a\sim\mathcal{D}_\theta}\left[\Phi(\widehat{\overline{\theta}}^\alpha;\theta^\alpha) \le \min\left\{\frac{1}{2}, \frac{3\log(1/\delta)}{n_\alpha}\right\} \mid n_\alpha\right] \le 1 - \frac{\delta}{4}.$$

Thus, using that $\Pr[\mathcal{E}_{\mathsf{Cher}}] \ge 1 - 2e^{-c_1pn}$,

$$\mathbb{E}_{\theta^\alpha\overset{\text{unif}}{\sim}\mathcal{S}^1}\Pr_{\mathcal{S}_a\sim\mathcal{D}_\theta}\left[\Phi(\widehat{\overline{\theta}}^\alpha;\theta^\alpha) \le \min\left\{\frac{1}{2}, \frac{3\log(1/\delta)}{2pn}\right\}\right] \le 1 - \frac{\delta}{4} - 2e^{-c_1pn}.$$

By considering the cases $\delta \le c_1pn$ and $d > c_1pn$ separately, some algebraic manipulations reveal that there exists a constant c, c' such that

$$\mathbb{E}_{\theta^\alpha\overset{\text{unif}}{\sim}\mathcal{S}^1}\Pr_{\mathcal{S}_a\sim\mathcal{D}_\theta}\left[\Phi(\widehat{\overline{\theta}}^\alpha;\theta^\alpha) \le c\min\left\{1, \frac{\log(1/\delta)}{pn}\right\}\right] \le 1 - c'\delta.$$

Thus, by Equations (29) and (28), we have

$$\Pr\left[\max_{a\in\{(0,-1),(0,1)\}}\min\{\mathbf{suf}_{\overline{f}_{\widehat{\overline{\theta}}};\overline{\mathcal{D}}_{\overline{\theta}}}(a;\overline{\mathcal{A}}_{\overline{w}}), \mathbf{cal}_{\overline{f}_{\widehat{\overline{\theta}}};\overline{\mathcal{D}}_{\overline{\theta}}}(a;\overline{\mathcal{A}}_{\overline{w}})\} \le c\min\left\{1, \frac{\log(1/\delta)}{pn}\right\}\right].$$

Part 2: Analysis of Least Squares Estimator As in the proof of Theorem C.1, we see that the least squares estimator $\widehat{f}_n$ takes the form $\overline{f}_{\widehat{\overline{w}}_{\text{LS}}}$.

$$\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\widehat{\overline{w}}_{\text{LS}}}) - \mathcal{L}^* = \mathbb{E}\|\widehat{\overline{w}}_{\text{LS}} - \overline{\theta}\|^2_{\frac{1}{16}\mathbb{E}[\overline{X}\overline{X}^\top]},$$

where again let $\|x\|_\Sigma^2 := x^\top\Sigma x$. Let X denote a random variable distributed uniformly on the sphere. By breaking the least squares estimator into components $\widehat{\overline{w}}_{\text{LS}} = (\theta^\alpha, \theta^\beta)$, we see that

1. Since $\overline{X}$ is either supported on the α component or the β component, θ^α and θ^β are the least-squares estimates on $\mathcal{S}_\alpha, \mathcal{S}_\beta$, respectively
2. Computing $\mathbb{E}[\overline{X}\overline{X}^\top] = \begin{bmatrix} p\mathbb{E}[XX^\top] & 0 \\ 0 & (1-p)\mathbb{E}[XX^\top] \end{bmatrix}$, we see that

$$\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\widehat{\overline{w}}_{\mathrm{LS}}}) - \mathcal{L}^* = p\mathbb{E}\|\theta^\alpha - \theta^\alpha\|^2_{\frac{1}{16}\mathbb{E}[XX^\top]} + (1-p)\mathbb{E}\|\theta^\beta - \theta^\beta\|^2_{\frac{1}{16}\mathbb{E}[XX^\top]}$$

With these two points in hand, we can use the analysis of the least squares estimator from Theorem C.1, conditioning on n_α and n_β, to find that with probability $1 - 2\delta - 2\exp(-c_3 \min\{n_\alpha, n_\beta\})$ that

$$\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\widehat{\overline{w}}_{\mathrm{LS}}}) - \mathcal{L}^* \le c_2 \left(\frac{p}{n_\alpha} + \frac{(1-p)}{n_\beta} \right) \log(1/\delta).$$

In particular, when $\mathcal{E}_{\mathsf{Cher}}$ holds, we have we have that with probability at least $1-2\delta-2\exp(-\frac{c_3}{2}\min\{pn, (1-p)n\}) =$, we have

$$\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\widehat{\overline{w}}_{\mathrm{LS}}}) - \mathcal{L}^* \le c_2 \left(\frac{p}{(pn/2)} + \frac{(1-p)}{((1-p)n/2)} \right) \log(1/\delta) = \frac{4c_2}{n} \log(1/\delta).$$

Finally, using the $\Pr[\mathcal{E}_{\mathsf{Cher}}] \ge 1 - 2e^{-\Omega(pn)}$, we conclude that for a new constant $c_2 \leftarrow 4c_2$, and new constant c_3 that

$$\overline{\mathcal{L}}_{\overline{\theta}}(\overline{f}_{\widehat{\overline{w}}_{\mathrm{LS}}}) - \mathcal{L}^* \le \frac{c_2 \log(1/\delta)}{n} \text{ with probability} 1 - 2\delta - 4e^{-c_3 pn}.$$

E Lower Bounds for Separation gap

Central to the proof is the following lemma, proved in Section E.1 below:

Lemma E.1. *Define the quantities*

$$Z_A := \mathbb{E}[(f^B)^2 \mid A], \quad q_A := \Pr[Y = 1 \mid A], \quad \overline{Z} := \mathbb{E}[(f^B)^2], \text{ and } \quad \overline{q} := \Pr[Y = 1]$$

Then, the following equalities hold.

$$\mathbb{E}[f^B \mid Y = 1, A] = \frac{Z_A}{q_A}, \quad \mathbb{E}[f^B \mid Y = 0, A] = \frac{q_A - Z_A}{1 - q_A},$$
$$\mathbb{E}[f^B \mid Y = 1] = \frac{\overline{Z}}{\overline{q}}, \quad \mathbb{E}[f^B \mid Y = 0] = \frac{\overline{q} - \overline{Z}}{1 - \overline{q}}.$$

We are now ready to finish the proof. By Lemma E.1 with $q_A = \Pr[Y = 1 \mid A]$,

$$\begin{aligned}
\mathbf{sep}_{f^B}(A) &= \mathbb{E}_A \left(q_A \cdot \left| \frac{\overline{Z}}{\overline{q}} - \frac{Z_A}{q_A} \right| + (1 - q_A) \left| \frac{q_A - Z_A}{1 - q_A} - \frac{\overline{q} - \overline{Z}}{1 - \overline{q}} \right| \right) \\
&= \mathbb{E}_A \left(\left| \frac{\overline{Z} q_A}{\overline{q}} - Z_A \right| + \left| q_A - Z_A - \left(\frac{1 - q_A}{1 - \overline{q}} \right) (\overline{q} - \overline{Z}) \right| \right) \\
&\ge \mathbb{E}_A \left| \frac{\overline{Z} q_A}{\overline{q}} - Z_A - \left(q_A - Z_A - \left(\frac{1 - q_A}{1 - \overline{q}} \right) (\overline{q} - \overline{Z}) \right) \right| \quad \text{(Reverse Triangle Inequality)} \\
&= \mathbb{E}_A \left| \frac{\overline{Z} q_A}{\overline{q}} + \left(\frac{1 - q_A}{1 - \overline{q}} \right) (\overline{q} - \overline{Z}) - q_A \right| \quad \text{(Cancelling } Z_A\text{)} \\
&= \mathbb{E}_A \left| \overline{Z} \left(\frac{q_A}{\overline{q}} - \frac{1 - q_A}{1 - \overline{q}} \right) + \left(\frac{1 - q_A}{1 - \overline{q}} \right) \overline{q} - q_A \right| \quad \text{(Grouping Terms).}
\end{aligned}$$

We further unpack

$$\begin{aligned}
&\mathbb{E}_A\left|\overline{Z}\left(\frac{q_A}{\overline{q}}-\frac{1-q_A}{1-\overline{q}}\right)+\left(\frac{1-q_A}{1-\overline{q}}\right)\overline{q}-q_A\right| \\
=&\mathbb{E}_A\left|(\overline{Z}-\overline{q})\left(\frac{q_A}{\overline{q}}-\frac{1-q_A}{1-\overline{q}}\right)\right| \\
=&\mathbb{E}_A\left|(\overline{Z}-\overline{q})\left(\frac{q_A(1-\overline{q})-(1-q_A)\overline{q}}{\overline{q}(1-\overline{q})}\right)\right| \\
=&\frac{|\overline{Z}-\overline{q}|}{\overline{q}(1-\overline{q})}\cdot\mathbb{E}_A|q_A-\overline{q}|,
\end{aligned}$$

Lastly, we find that

$$\begin{aligned}
|\overline{Z}-\overline{q}| &= |\mathbb{E}[(f^B)^2]-\Pr[Y=1]| \;= |\mathbb{E}[(f^B)^2]-\mathbb{E}[f^B]| \\
&= |\mathbb{E}[(f^B)(f^B-1)]| \;= \mathbb{E}[(f^B)(1-f^B)] \quad \text{since } f^B\in[0,1] \\
&= \mathbb{E}[(\mathbb{E}[Y|X,A])(1-\mathbb{E}[Y|X,A])] \;= \mathbb{E}[\mathrm{Var}[Y|X,A]].
\end{aligned}$$

Moreover, $\overline{q}(1-\overline{q}) = \mathbb{E}[Y](1-\mathbb{E}[Y]) = \mathrm{Var}[Y]$. Thus

$$\mathbf{sep}_{f^B}(A) \geq \frac{|\overline{Z}-\overline{q}|}{\overline{q}(1-\overline{q})}\cdot\mathbb{E}_A|q_A-\overline{q}| \geq \frac{\mathbb{E}[\mathrm{Var}[Y|X,A]]}{\mathrm{Var}[Y]}\cdot\mathbb{E}_A|q_A-\overline{q}|.$$

E.1 Proof of Lemma E.1

For ease of notation, we write $f = f^B$. Observe then that by the definition of the Bayes classifer,

$$\Pr[Y=1 \mid f(X,A)=\tau, A] = \tau$$

First we compute the conditional densities by Bayes rule:

$$\begin{aligned}
\Pr[f(X,A)=\tau \mid Y=1,A] &= \frac{\Pr[Y=1 \mid f(X,A)=\tau,A]\Pr(f(X,A)=\tau \mid A)}{\Pr[Y=1 \mid A]} \\
&= \frac{\tau\Pr(f(X,A)=\tau \mid A)}{\Pr[Y=1 \mid A]}, \\
\Pr[f(X,A)=\tau \mid Y=0,A] &= \frac{\Pr[Y=0 \mid f(X,A)=\tau,A]\Pr(f(X,A)=\tau \mid A)}{\Pr[Y=0 \mid A]} \\
&= \frac{(1-\tau)\Pr(f(X,A)=\tau \mid A)}{\Pr[Y=0 \mid A]}.
\end{aligned}$$

Integrating these to compute the relevant expectations, we have

$$\begin{aligned}
\mathbb{E}[f \mid Y=1, A] &= \int \tau \Pr[f(X,A)=\tau \mid Y=1, A] d\tau \\
&= \int \frac{\tau^2 Y}{\Pr[Y=1 \mid A]} \Pr(f(X,A)=\tau \mid A) d\tau \\
&= \frac{\mathbb{E}[f^2 \mid A]}{\Pr[Y=1 \mid A]}, \\
\mathbb{E}[f \mid Y=0, A] &= \int \tau \Pr[f(X,A)=\tau \mid Y=0, A] d\tau \\
&= \int \frac{\tau(1-\tau)}{\Pr[Y=0 \mid A]} \Pr[f(X,A)=\tau \mid A] d\tau \\
&= \frac{\Pr[Y=1 \mid A] - \mathbb{E}[f^2 \mid A]}{\Pr[Y=0 \mid A]}, \qquad \text{since } \mathbb{E}[f^B \mid A] = \Pr[Y=1 \mid A].
\end{aligned}$$

In summary, we have

$$\begin{aligned}
\mathbb{E}[f \mid Y=1, A] &= \frac{\mathbb{E}[f^2 \mid A]}{\Pr[Y=1 \mid A]} = \frac{Z_A}{q_A}. \\
\mathbb{E}[f \mid Y=0, A] &= \frac{(\Pr[Y=1 \mid A] - \mathbb{E}[f^2 \mid A])}{\Pr[Y=0 \mid A]} = \frac{q_A - Z_A}{1-q_A}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathbb{E}[f \mid Y=1] &= \frac{\mathbb{E}[f^2]}{\Pr[Y=1]} = \frac{\overline{Z}}{\overline{q}}. \\
\mathbb{E}[f \mid Y=0] &= \frac{(\Pr[Y=1] - \mathbb{E}[f^2])}{\Pr[Y=0]} = \frac{\overline{q} - \overline{Z}}{1-\overline{q}}.
\end{aligned}$$

E.2 Proof of Corollary 2.6

By the reverse triangle inequality, the definition of separation, and Jensen's inequality, we bound

$$\begin{aligned}
\mathbf{sep}_f(A) &:= \mathbb{E}_{Y,A}[|\mathbb{E}[f(X,A) \mid Y, A] - \mathbb{E}[f(X,A) \mid Y]|] \\
&\geq \mathbb{E}_{Y,A}[|\mathbb{E}[f^B(X) \mid Y, A] - \mathbb{E}[f^B(X) \mid Y]|] - \mathbb{E}_{Y,A}[|\mathbb{E}[f - f^B \mid Y, A] - \mathbb{E}[f - f^B(X) \mid Y]|] \\
&= \mathbf{sep}_{f^B}(A) - \mathbb{E}_{Y,A}[|\mathbb{E}[f^B(X) \mid Y, A] - \mathbb{E}[f^B(X) \mid Y]|] - \mathbb{E}_{Y,A}[|\mathbb{E}[f - f^B \mid Y, A] \\
&\geq \mathbf{sep}_{f^B}(A) - 2\mathbb{E}_{Y,A}[|f - f^B|].
\end{aligned}$$

By Proposition 2.5, we have $\mathbf{sep}_{f^B} \geq C_{f_B} \mathbb{E}[|\overline{q} - q_A|]$. Moreover, we have

$$\mathbb{E}_{Y,A}[|f - f^B|] \leq \sqrt{\mathbb{E}_{Y,A}[|f - f^B|^2]} \leq \sqrt{\frac{\mathcal{L}(f) - \mathcal{L}(f^B)}{\kappa}},$$

where the first inequality is Jensen's inequality, and the second using κ-strong convexity of $\mathcal{L}$ as in the proof of Theorem 2.3.

F Lower Bound for Independence Gap

We define the independence gap of a score f with respect to a group attribute A as

$$\mathbf{ind}_f(A) = \mathbb{E}[|\mathbb{E}[f|A] - \mathbb{E}[f]|] \tag{31}$$

Note that $\mathbf{ind}_f(A) = 0$ if and only if f and A are conditionally mean independent, which is implied by (but does not imply) statistical independence. The following proposition shows that any score with a small excess risk must have a large independence gap if the base rate $\Pr[Y = 1]$ differs by group.

Proposition F.1 (Independence of Unconstrained Learning)**.** *Let $\mathcal{L}$ be the risk associated with a loss function $\ell(\cdot, \cdot)$ satisfying Assumption 1 with parameter $\kappa > 0$. Denote $\overline{q} := \Pr[Y = 1]$, and $q_A := \Pr[Y = 1|A]$ for a group attribute A. Then, for any score $\hat{f} \in \mathcal{F}$,*

$$\mathbf{ind}_{\hat{f}}(A) \geq \mathbb{E}[|q_A - \overline{q}|] - 2\sqrt{\frac{\mathcal{L}(\hat{f}) - \mathcal{L}(f^B)}{\kappa}}$$

Proof. Observe that for the calibrated Bayes score f^B, $\mathbb{E}[f^B|A] = q_A := \Pr[Y = 1|A]$, whereas $\mathbb{E}[f^B] = \overline{q} = \Pr[Y = 1]$. Hence

$$\mathbf{ind}_{f^B}(A) = \mathbb{E}[|\mathbb{E}[f^B|A] - \mathbb{E}[f^B]|] = \mathbb{E}[|q_A - \overline{q}|]. \tag{32}$$

We now lower bound the $\mathbf{ind}_f(A)$ for arbitrary f. The remainder of the proof follows along the lines of Corollary 2.6. By the reverse triangle inequality, the definition of separation, and Jensen's inequality, we bound

$$\begin{aligned}
\mathbf{ind}_f(A) &:= \mathbb{E}_A[|\mathbb{E}[f|A] - \mathbb{E}[f]|] \\
&\geq \mathbb{E}_A[|\mathbb{E}[f^B(X) \mid A] - \mathbb{E}[f^B(X)]|] - \mathbb{E}_A[|\mathbb{E}[f - f^B \mid A] - \mathbb{E}[f - f^B(X)]|] \\
&= \mathbf{ind}_{f^B}(A) - \mathbb{E}_A[|\mathbb{E}[f - f^B \mid A] - \mathbb{E}[f - f^B(X)]|] \\
&\geq \mathbf{ind}_{f^B}(A) - 2\mathbb{E}_{Y,A}[|f - f^B|].
\end{aligned}$$

By (32), we have $\mathbf{ind}_{f^B}(A) \geq \mathbb{E}[|\overline{q} - q_A|]$. Moreover, we have

$$\mathbb{E}_{Y,A}[|f - f^B|] \leq \sqrt{\mathbb{E}_{Y,A}[|f - f^B|^2]} \leq \sqrt{\frac{\mathcal{L}(f) - \mathcal{L}(f^B)}{\kappa}},$$

where the first inequality is Jensen's inequality, and the second using κ-strong convexity of $\mathcal{L}$ as in the proof of Theorem 2.3.

□

G Addendum to experiments

G.1 Empirical estimate of $\mathbf{suf}_f(A)$

We estimate $\mathbf{suf}_f$ from the test set and the scores for this test set, that is $\{x_i, y_i, a_i, f(x_i)\}_{i=1}^n$. We divide the scores into deciles, that is, $B = 10$ equally spaced intervals on $[0, 1]$. For any score value

$f \in [0,1]$, let $d(f)$ denote the corresponding decile for f. We estimate $\mathbb{E}[Y|f]$ as the average rate of positive outcomes in the corresponding decile of the score f, i.e.

$$\hat{g}(f) = \frac{1}{N} \sum_{i=1}^{n} y_i \mathbf{1}\{f(x_i) \in d(f)\},$$

where $N = \sum_{i=1}^{n} \mathbf{1}\{f(x_i) \in d(f)\}$. For any group a and score f, we estimate $\approx \mathbb{E}[Y|f, A = a]$ as the average rate of positive outcomes in the corresponding decile of the score f in group a, i.e.

$$\hat{g}(f, a) = \frac{1}{M} \sum_{i=1}^{n} y_i \mathbf{1}\{f(x_i) \in d(f) \cap a_i = a\},$$

where $M = \sum_{i=1}^{n} \mathbf{1}\{f(x_i) \in d(f) \cap a_i = a\}$. $\widehat{\mathbf{suf}}_f$ is then computed as the sample average $\frac{1}{n} \sum_{i=1}^{n} |\hat{g}(f(x_i), a_i) - \hat{g}(f(x_i))|$. In general, the value of the estimate does vary with the chosen number of intervals B. We find that on the Adult dataset, for example, the choice $B = 10$ results in bins that all have a suitable number of samples, and hence provides an adequate estimate of the expected sufficiency gap for experimental purposes. We leave the statistical properties of this estimator, such as the ramifications of different B, to future work.

G.2 Broward dataset

Compared to the Adult dataset, the Broward dataset contains about 6 times fewer training and testing examples; as expected, our estimates of the sufficiency gap are noisier. Figure 7 shows that the score obtained from empirical risk minimization with the logistic loss is largely calibrated with respect to gender, race and age across different scores, barring score buckets where there was grossly insufficient data to estimate the rate of positive outcomes.

In Figure 8, we show the calibration and separation error of the logistic regression model as we use more training examples. The average and standard deviation (indicated as confidence intervals) are computed over 20 random draws of training examples. To deal with insufficient data in certain score deciles, we instead estimated the sufficiency gap using 8 score buckets based on quantiles. For race, the sufficiency gap is decreasing with the number of samples, while the separation gap does not decrease, stabilizing at a value of 0.05. For gender, the sufficiency gap is relatively small to begin with (0.03) and appears to remain at the same level with more samples.

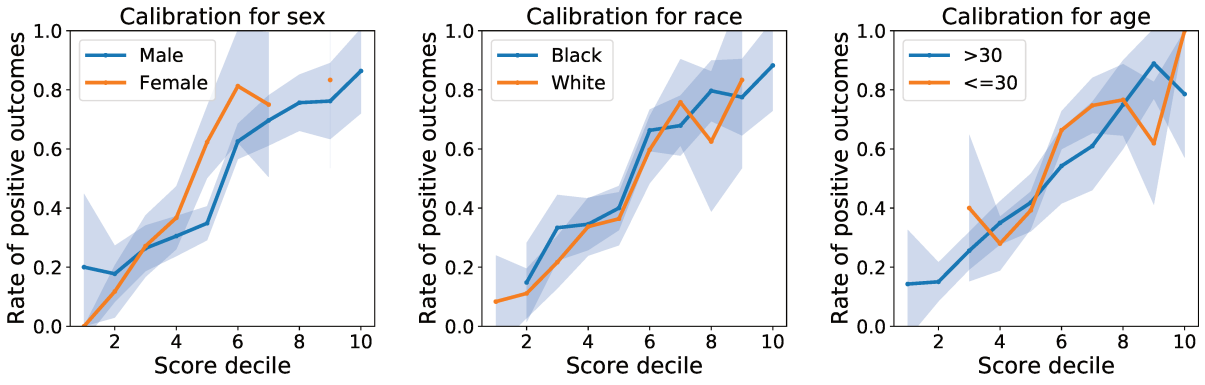

Figure 7: Calibration plots with respect to group attributes for the Broward dataset. Missing datapoints are where the score decile bucket contains fewer than two individuals.

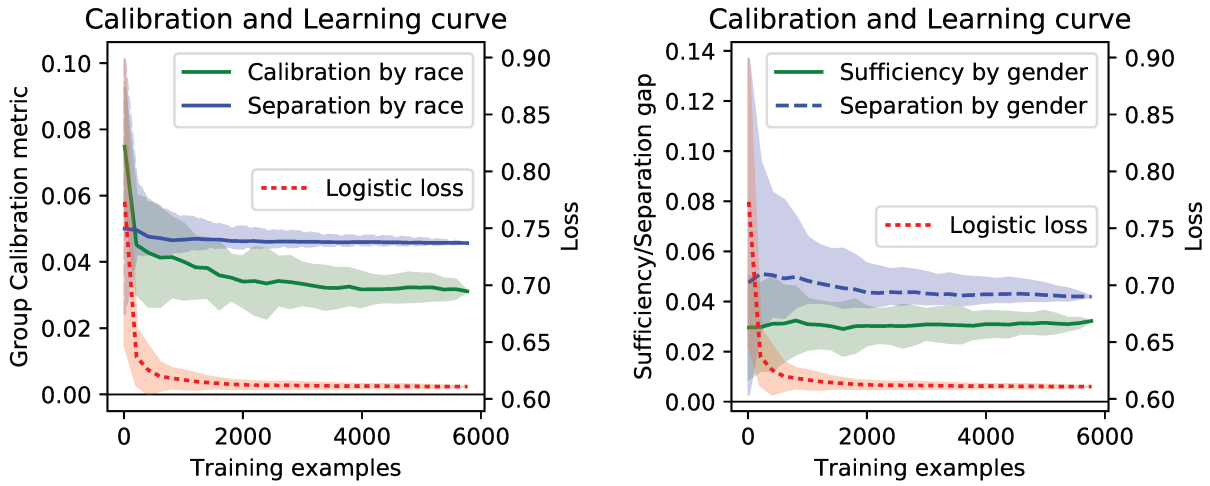

Figure 8: Sufficiency, Separation, and Loss vs. Number of training examples for the Broward dataset

G.3 Simultaneous calibration with respect to multiple, rich group attributes

In this section, we present additional results on the Adult dataset. Specifically, we show the calibration plots for the score obtained from logistic regression, with respect to a variety of group attributes, including 'fabricated' group attributes that are a combination of two features. For numerical features (e.g. Age), we split the data into 2 groups according to an arbitrary threshold (e.g. above 40 years old and below 40 years old) and compute calibration with respect to those groups. For categorical features, we only visualize the calibration of the top three most populous groups, for clarity. This is shown in Figure 9. Note that by Theorem 2.3, groups that have small mass have a worse bound on calibration.

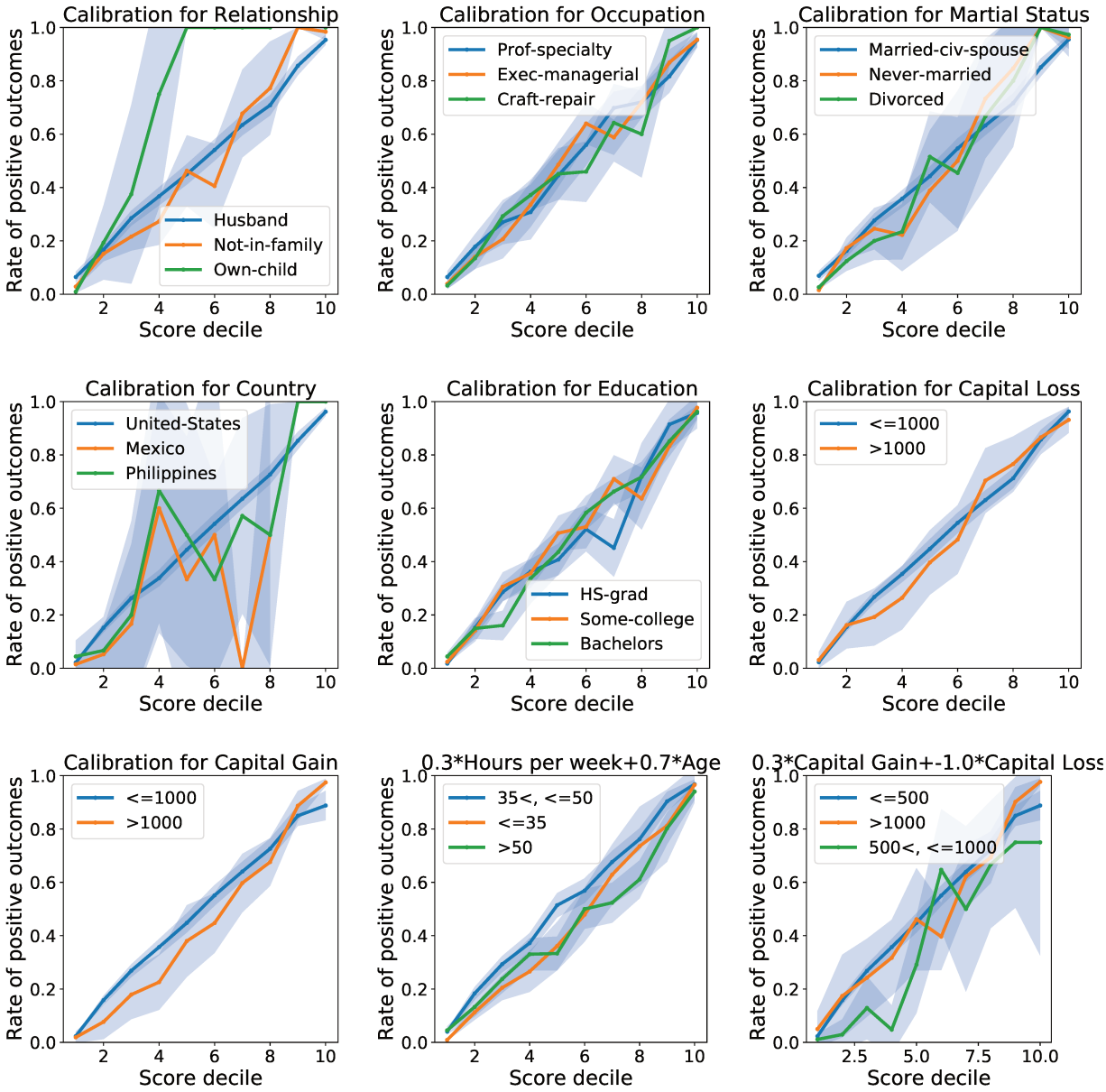

Figure 9: Calibration plots with respect to other features and combinations of features for the Adult dataset