Learning Non-Discriminatory Predictors

Blake Woodworth   BLAKE@TTIC.EDU
Suriya Gunasekar   SUIRYA@TTIC.EDU
Mesrob I. Ohannessian   MESROB@TTIC.EDU
Nathan Srebro   NATI@TTIC.EDU
Department of Computer Science, Toyota Technological Institute at Chicago

Abstract

We consider learning a predictor which is non-discriminatory with respect to a "protected attribute" according to the notion of "equalized odds" proposed by Hardt et al. (2016). We study the problem of learning such a non-discriminatory predictor from a finite training set, both statistically and computationally. We show that a post-hoc correction approach, as suggested by Hardt et al, can be highly suboptimal, present a nearly-optimal statistical procedure, argue that the associated computational problem is intractable, and suggest a second moment relaxation of the non-discrimination definition for which learning is tractable.

1. Introduction

Machine learning algorithms are increasingly deployed in important decision making tasks that affect people's lives significantly. These tools already appear in domains such as lending, policing, criminal sentencing, and targeted service offerings. In many of these domains, it is morally and legally undesirable to discriminate based on certain "protected attributes" such as race and gender. Even in seemingly innocent applications, such as ad placement and product recommendations, such discrimination might be illegal or detrimental. Consequently, there has been abundant public, academic and technical interest in notions of non-discrimination and fairness, and achieving "equal opportunity by design" is a major United States national Big Data challenge, White House (2016).

We consider non-discrimination in supervised learning where the goal is to learn a (potentially randomized) predictor $h(X)$ or[1] $h(X, A)$ for a target quantity Y using features X and a protected attribute A, while ensuring non-discrimination with respect to A. As an illustrative example, consider a financial institution that wants to predict whether a particular individual will pay back a loan or not, corresponding to $Y = 1$ and $Y = 0$, respectively. The features X could include financial as well as other information, e.g. about education, driving, and housing history, languages spoken, and the number of members in the household, all of which have a potential of being used inappropriately as a surrogate for a protected attribute A, such as gender or race. It is important that the predictor for loan repayment not be even implicitly discriminatory with respect to A.

Recent work has addressed the issue of defining what it means to be non-discriminatory—both in the context of supervised learning e.g. Dwork et al. (2012); Pedreshi et al. (2008); Feldman et al. (2015), and otherwise e.g. Joseph et al. (2016b,a). The particular notion of non-discrimination we consider here is "equalized odds", recently presented and studied by Hardt et al. (2016):

---

1. See Hardt et al. (2016) for a discussion on why it might be necessary for a non-discriminatory predictor to use A

Definition 1 (Equalized odds) *A possibly randomized predictor $\widehat{Y} = h(X, A)$ for target Y is non-discriminatory with respect to a protected attribute A if $\widehat{Y}$ is independent of A conditioned on Y.*

Informally, we require that even when the correct label Y provides information about the protected attribute A, if we already know Y, the prediction $\widehat{Y}$ does not provide any *additional* information about A. The definition can also be motivated in terms incentive structure and of moving the burden of uncertainty from the protected population to the decision maker. See Hardt et al. (2016) for further discussion of the definition, its implications, and comparisons to alternative notions.

In a binary prediction task with binary protected attribute, i.e. $\widehat{Y}, A, Y \in \{0, 1\}$, Definition 1 can be qualified in terms of true and false positive rates. Denote the group-conditional true and false positive rates as,

$$\gamma_{ya}(\widehat{Y}) := \mathbb{P}(\widehat{Y} = 1 \mid Y = y, A = a), \tag{1}$$

Then Definition 1 is equivalent to requiring that the class conditional true and false positive rates agree across different groups (different values of A):

$$\gamma_{00}(\widehat{Y}) = \gamma_{01}(\widehat{Y}) \qquad \text{and} \qquad \gamma_{10}(\widehat{Y}) = \gamma_{11}(\widehat{Y}) \tag{2}$$

Returning to the loan example, this definition requires that the percentage of men who are wrongly denied loans even though they would have paid it back must match the corresponding percentage for women, and similarly the percentage of men who are wrongly given loans that they will not pay back must also match the corresponding percentage of women. This does *not* however require that the same percentage of male and female applicants will receive loans. For instance, if women pay back loans with *truly* higher frequency than men, then the predictor would be allowed to deny loans to men more often than women.

While Hardt et al. focused on the notion itself and how it behaves on the population, in this work we tackle the problem of how to learn a good non-discriminatory predictor (i.e. satisfying the equalized odds) from a finite training set. We examine this both from a statistical perspective of how to best obtain a predictor from finite data that would be as accurate and non-discriminatory as possible on the population, and from a computational perspective.

One possible approach to learning a non-discriminative predictor is the *post hoc correction* proposed by Hardt et al. (2016): first learn a good, possibly discriminatory predictor. Afterwards, this predictor is "corrected" by taking into account A in order to make the predictor non-discriminatory. When Y is binary and the predictors $\widehat{Y}$ are real-valued, they show that the unconstrained Bayes optimal least-square regressor can be post hoc corrected to the optimal predictor with respect to the 0-1 loss. In Section 2, we consider more carefully the limitations of such a post hoc procedure. In particular, we show that this approach can fail for the 0-1 and hinge losses, even if the Bayes optimal predictor with respect to those losses is learned in the first step. We also show that even when minimizing the squared loss, the approach can fail once the hypothesis class is constrained, as is essential when learning from finite data. From this, we conclude that post hoc correction is not sufficient, and that it is necessary to directly incorporate non-discrimination into the learning process.

Turning to learning from finite data, we cannot hope to ensure exact non-discrimination on the population. To this end, in Section 3 we define a notion of approximate non-discrimination, motivate it, and explore its limits by analyzing the statistical problem of detecting whether or not a predictor is at least α-discriminatory.

We then turn to the main statistical question: given a finite training set, how can we best learn a predictor that is ensured to be as non-discriminatory as possible (on the population) and competes (in terms of its population loss) with the best non-discriminatory predictor in some given hypothesis class (this is essentially an extension of the notion of agnostic PAC learning with a non-discrimination constraint). In Section 4 we show that an ERM-type procedure, minimizing the training error subject to an empirical non-discrimination constraint, is statistically sub-optimal, and instead we present a statistically optimal (up to constant factors) two-step learning procedure for non-discriminatory binary classification.

Unfortunately, learning a non-discriminatory binary classifier is computationally hard, which we prove in Section 5. In order to allow tractable training, in Section 6, we present a relaxation of equalized odds, based only on a second-moment condition instead of full conditional independence. We show that under this second moment notion of non-discrimination it is computationally tractable to learn a nearly optimal non-discriminatory linear predictor with respect to a convex loss.

2. Sub-optimality of post hoc correction

When the protected attribute A and the target Y are both binary, the post hoc correction algorithm proposed by Hardt et al. (2016) can be applied to a binary or real-valued predictor $\widehat{Y} \in \mathcal{H}$, deriving a randomized binary predictor that is non-discriminatory. The algorithm is convenient because it requires access only to the joint distribution over $(\widehat{Y}, A, Y)$ and does not use the features X, thus it can be applied retroactively to an already trained predictor. Such predictors are formulated using the notion of a derived predictor:

Definition 2 (Definition 4.1 in Hardt et al. (2016)) *A predictor $\widetilde{Y}$ is **derived** from a random variable R and protected attribute A if it is a possibly randomized function of (R, A) alone. In particular, $\widetilde{Y}$ is independent of X conditioned on (R, A).*

For binary classification, the optimal post hoc correction $\widetilde{Y}$ for a binary or real valued predictor $\widehat{Y} \in \mathbb{R}$ is simply the non-discriminatory, derived, binary predictor that minimizes the expectation of a loss ℓ over binary variables (Hardt et al., 2016):

$$\widetilde{Y} = \underset{f:\mathbb{R}\times\{0,1\}\mapsto\{0,1\}}{\text{argmin}} \mathbb{E}\,\ell\left(f(\widehat{Y}, A), Y\right)$$

$$\text{s.t.} \quad \gamma_{y0}(f) = \gamma_{y1}(f) \qquad \forall y = \{0, 1\} \tag{3}$$

Two notable features of the corrected predictor $\widetilde{Y}$ are that **a)** it is not constrained to any particular hypothesis class, and **b)** it may be a random function of $\widehat{Y}$ and A; indeed for many distributions and hypothesis classes there may not even exist a non-constant, deterministic, non-discriminatory predictor. Nevertheless, $\widetilde{Y}$ does indirectly depend on the hypothesis class $\mathcal{H}$ from which $\widehat{Y}$ was learned and possibly a different loss over real valued variables used in training of $\widehat{Y}$.

We are interested in comparing the optimality of $\widetilde{Y}$ from post hoc correction to the following Y^* which is the optimal non-discriminatory predictor in a hypothesis class $\mathcal{H}$ under consideration:

$$Y^* = \underset{h\in\mathcal{H}}{\text{argmin}}\, \mathbb{E}\,\ell\left(h(X, A), Y\right) \qquad \text{s.t.} \quad \gamma_{y0}(h) = \gamma_{y1}(h) \qquad \forall y = \{0, 1\} \tag{4}$$

Ideally, the expected loss of $\widetilde{Y}$ would compare favorably against that of Y^*. Indeed, Hardt et al. (2016) show that when the target Y is binary, if we can first find a predictor $\widehat{Y}$ that is exactly or

nearly Bayes optimal for the squared loss over an unconstrained hypothesis class, then applying the post hoc correction (3) using the 0-1 loss (i.e. with $\ell = \ell^{01}$ in (3)) to $\widehat{Y}$ will yield a predictor $\widetilde{Y}$ that is non-discriminatory and has loss no worse than Y^*. This statement can be extended to the case of first finding the optimal *unconstrained* predictor with resepect to any *strictly convex* loss, and then using the post hoc correction (3) with the 0-1 loss.
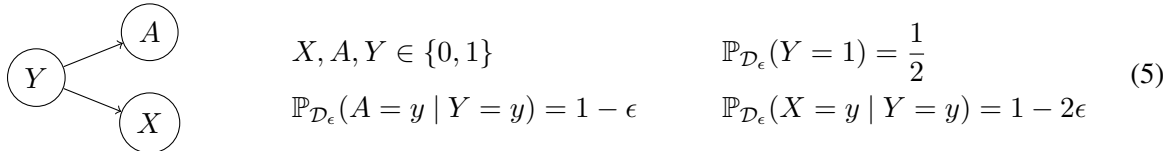
Nevertheless, from a practical perspective this approach is very unsatisfying. First, for general distributions, it is impossible to learn the Bayes optimal predictor from finite samples of data. Also, as we will show, the post hoc correction of even the optimal unconstrained predictor with respect to the 0-1 (non-convex) or even hinge (non-strict but convex) losses can have much worse performance than the best non-discriminatory predictor. Moreover, if the hypothesis class is restricted there can also be a gap between the post hoc correction of the optimal predictor in the hypothesis class and the best non-discriminatory predictor, even when optimizing a strictly convex loss function.

In the following example, we see that when the loss function is not strictly convex, the post hoc correction of even the unconstrained Bayes optimal predictor can have poor accuracy:

Example 1 *When the hypothesis class is unconstrained, for any $\epsilon \in (0, 1/4)$ there exists a distribution $\mathcal{D}_\epsilon$ such that a) the optimal non-discriminatory predictor Y^* with respect to the 0-1 loss has loss at most 2ϵ but b) for unrestricted Bayes optimal predictor $\widehat{Y}$ trained on 0-1 loss, the post hoc correction of $\widehat{Y}$ with 0-1 loss returns a predictor $\widetilde{Y}$ with loss at least 0.5.*

A similar statement can also be made about predictors trained on hinge loss. For an unconstrained hypothesis class, for any $\epsilon \in (0, 1/4)$ and the same distribution $\mathcal{D}_\epsilon$, a) the optimal non-discriminatory predictor Y^* with respect to the hinge loss has loss at most 4ϵ but b) the post hoc correction of the Bayes optimal unrestricted predictor trained on hinge loss has loss 1.

We construct $\mathcal{D}_\epsilon$ as follows:

$$
\begin{aligned}
&X, A, Y \in \{0, 1\} \qquad\qquad && \mathbb{P}_{\mathcal{D}_\epsilon}(Y = 1) = \frac{1}{2} \\
&\mathbb{P}_{\mathcal{D}_\epsilon}(A = y \mid Y = y) = 1 - \epsilon \qquad\qquad && \mathbb{P}_{\mathcal{D}_\epsilon}(X = y \mid Y = y) = 1 - 2\epsilon
\end{aligned}
\tag{5}
$$

Both X and A are highly predictive of Y, but A is slightly more so. Therefore, minimizing either the 0-1 or the hinge loss, without regard for non-discrimination, returns $\widehat{Y} = A$ and ignores X entirely. Consequently, $\gamma_{y1}(\widehat{Y}) = 1$ and $\gamma_{y0}(\widehat{Y}) = 0 \neq \gamma_{y1}(\widehat{Y})$ so the Bayes optimal predictor is discriminatory, and the post hoc correction, which is required to be non-discriminatory and derived from $\widehat{Y} = A$, is forced to return a constant predictor even though returning $Y^* = X$ would be accurate and non-discriminatory. A more detailed proof is included in Appendix A.1

In the second example, we show that when the hypothesis class is restricted, the correction of the optimal regressor in the class can yield a suboptimal classifier, even with squared loss.

Example 2 *Let $\mathcal{H}$ be the class of linear predictors with L^1 norm at most $\frac{1}{2} - 2\epsilon$, for some $\epsilon \in (2/25, 1/4)$. There exists a distribution $\mathcal{D}_\epsilon$ such that a) the optimal non-discriminatory predictor in $\mathcal{H}$ with respect to the squared loss has square loss at most $\frac{1}{16} + \frac{3\epsilon}{2} + 3\epsilon^2$, but b) the post hoc correction of the Bayes optimal square loss regressor in $\mathcal{H}$ returns a constant predictor which has (trivial) square loss of $1/4$.*

Similarly, for the class $\mathcal{H}$ of sparse linear predictors, for any $\epsilon \in (0, 1/4)$, there exists a distribution $\mathcal{D}_\epsilon$ such that a) the optimal non-discriminatory predictor in $\mathcal{H}$ with respect to the squared

loss has square loss at most $2\epsilon - 4\epsilon^2$, but b) the post hoc correction of the Bayes optimal squared loss regressor in $\mathcal{H}$ again returns a constant predictor which has (trivial) square loss of $1/4$.

The distribution $\mathcal{D}_\epsilon$ is the same as was defined in (5). Again, A is slightly more predictive of Y than X, and since the sparsity or the sparsity surrogate L^1 norm of the predictor is constrained by the hypothesis class, the Bayes optimal predictor chooses to use just the feature A and ignore X. Consequently, the optimal predictor is extremely discriminatory, and the post hoc correction algorithm will return a highly sub-optimal constant predictor which performs no better than chance. Details of the proof are deferred to Appendix A.2

From these examples, it is clear that simply finding the optimal predictor with respect to a particular loss function and hypothesis class and correcting it post hoc can perfom very poorly. We conclude that in order to learn a predictor that is simultaneously accurate *and* non-discriminatory in the general case, it is essential to account for non-discrimination during the learning process.

3. Detecting Discrimination in Binary Predictors

In the following sections, we look at tools for integrating non-discrimination into the supervised learning framework. In formulating algorithms for learning non-discriminatory predictors, it is important to consider the non-asymptotic behavior under finite samples. Towards this, one of the first issues to be addressed is that, using finite samples it is not feasible to ensure, or even verify if a predictor $\widehat{Y}$ satisfies the non-discrimination criterion in Definition 1. This necessitates defining a notion of approximate non-discrimination which can be computed using finite samples and which asymptotically generalizes to the equalized odds criteria in the population.

Let us consider the task of binary classification, where both $A, Y \in \{0, 1\}$ and the predictors $\widehat{Y}$ output values in $\{0, 1\}$. Recall the definition of the population group-conditional true and false positive rates $\gamma_{ya}(\widehat{Y}) = \mathbb{P}(\widehat{Y} = 1 \mid Y = y, A = a)$ and the fact that non-discrimination is equivalent to satisfying $\gamma_{00} = \gamma_{01}$ and $\gamma_{10} = \gamma_{11}$.

For a set of of n i.i.d. samples, $S = \{(x_i, a_i, y_i)\}_{i=1}^n \sim \mathbb{P}^n(X, A, Y)$, the sample analogue of γ_{ya} is defined as follows,

$$\gamma_{ya}^S(\widehat{Y}) = \frac{1}{n_{ya}^S} \sum_{i=1}^n \widehat{Y}(x_i, a_i) \mathbf{1}(y_i = y, a_i = a), \text{ where } n_{ya}^S = \sum_{i=1}^n \mathbf{1}(y_i = y, a_i = a). \quad (6)$$

To ensure non-discrimination, we could possibly require $\gamma_{y0}^S = \gamma_{y1}^S$ on a large enough sample S, however this not ideal for two reasons. First, even when $\gamma_{y0}^S = \gamma_{y1}^S$ on S, this almost certainly does not ensure that $\gamma_{y0} = \gamma_{y1}$ on the population. For this same reason, it is impossible to be certain that a given predictor is non-discriminatory on the population. Moreover, if $n_{y0}^S \neq n_{y1}^S$, it is typically not feasible to match $\gamma_{y0}^S = \gamma_{y1}^S$ for non-trivial predictors, e.g. if $n_{y0}^S = 2$ and $n_{y1}^S = 3$, then $\gamma_{y0}^S \in \{0, \frac{1}{2}, 1\}$ but $\gamma_{y1}^S \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$, thus the only predictors with $\gamma_{y1}^S = \gamma_{y0}^S$ would be the ones which are constant conditioned on Y, i.e. the constant predictors $\widehat{Y} = 0$ and $\widehat{Y} = 1$, or the perfect predictor $\widehat{Y} = Y$.

For these reasons, we define the following notion of approximate non-discrimination, which *is* possible to ensure on a sample and, when it holds on a sample, generalizes to the population.

Definition 3 (α-discrimination) *A possibly randomized binary predictor $\widehat{Y}$ is α-discriminatory with respect to a binary protected attribute A on the population or on a sample S if, respectively,*

$$\Gamma(\widehat{Y}) := \max_{y \in \{0,1\}} \left| \gamma_{y0}(\widehat{Y}) - \gamma_{y1}(\widehat{Y}) \right| \leq \alpha \quad or \quad \Gamma^S(\widehat{Y}) := \max_{y \in \{0,1\}} \left| \gamma_{y0}^S(\widehat{Y}) - \gamma_{y1}^S(\widehat{Y}) \right| \leq \alpha. \quad (7)$$

The decision to define approximate non-discrimination in terms of *conditional* rather than joint probabilities is important, particularly in the case that the a, y pairs occur with widely varying frequencies. For example, if approximate non-discrimination were defined in terms of the joint probabilities $P(\widehat{Y} = \widehat{y}, A = a, Y = y)$ and if $P(A = 0, Y = 1) = \alpha/10$, then a predictor could be "α-discriminatory" all while being arbitrarily unfair towards the $A = 0, Y = 1$ population. This issue does not arise when using Definition 3 and it incentivizes collection of sufficient data for minority groups to ensure non-discrimination.

For Definition 3, we propose a simple statistical test to test the hypothesis that a given predictor $\widehat{Y}$ is at most α-discriminatory on the population for some $\alpha > 0$. Let $S = \{(x_i, a_i, y_i)\}_{i=1}^n \sim \mathbb{P}^n(X, Y, A)$ denote a set of n i.i.d. samples, and for $y, a \in \{0, 1\}$, let $\mathrm{P}_{ya} = \mathbb{P}(Y = y, A = a)$. We propose the following test for detecting α-discrimination:

$$T\left(\widehat{Y}, S, \alpha\right) = \mathbf{1}\left(\Gamma^S(\widehat{Y}) > \alpha\right) \quad (8)$$

Lemma 4 *Given n i.i.d. samples S, $\forall \alpha \in (0, 1), \delta \in (0, 1/2)$, if $n > \frac{16 \log 32/\delta}{\alpha^2 \min_{ya} \mathrm{P}_{ya}}$, then with probability greater than $1 - \delta$, T satisfies,*

$$T\left(\widehat{Y}, S, \frac{\alpha}{2}\right) = \begin{cases} 0 & \text{if } \widehat{Y} \text{ is 0-discriminatory on population} \\ 1 & \text{if } \widehat{Y} \text{ is at least } \alpha\text{-discriminatory on population.} \end{cases}$$

The proof is based on the following concentration results for Γ^S and is provided in Appendix B.

Lemma 5 *For $\delta \in (0, 1/2)$ and a binary predictor h, if $n > \frac{8 \log 8/\delta}{\min_{ya} \mathrm{P}_{ya}}$, then*

$$\mathbb{P}\left(\left|\Gamma(h) - \Gamma^S(h)\right| > 2 \max_{ya} \sqrt{\frac{\log 16/\delta}{n \mathrm{P}_{ya}}}\right) \leq \delta.$$

4. Learning Optimal Non-discriminatory Binary Predictors

In Example 1, we saw that even though an almost perfect non-discriminatory predictor exists within the hypothesis class, if we ignore non-discrimination in training with 0-1 loss, then optimal post hoc correction of even the Bayes optimal predictor using (3) yields a poor predictor with no better than chance accuracy. Thus, to find a predictor that is both nearly non-discriminatory *and* has nearly optimal loss for general hypothesis classes, it is necessary to incorporate non-discrimination into the learning process. For a hypothesis class $\mathcal{H}$, we would ideally like to find the optimal non-discriminatory predictor:

$$Y^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \, \mathbb{E}\ell\left(h(X, A), Y\right) \quad \text{s.t.} \quad \gamma_{y0}(h) = \gamma_{y1}(h) \quad \forall y = \{0, 1\}. \quad (9)$$

6

However, as discussed in Section 3, it is impossible to learn 0-discriminatory predictors from finite samples. In this section, we address following question: *given n i.i.d samples S, what level of approximate non-discrimination and accuracy is it possible to ensure in a learned predictor?*

We propose a two step framework for learning a non-discriminatory binary predictor that minimizes the expected 0-1 loss $\mathcal{L}(\widehat{Y}) = \mathbb{E}\,\ell^{01}(\widehat{Y}, Y) = \mathbb{E}\,\mathbf{1}(\widehat{Y} \neq Y)$ over a binary hypothesis class $\mathcal{H} = \{h : \mathcal{X} \to \{0, 1\}\}$. Broadly, the two-step framework is as follows:

1. **Non-discrimination in training**: Estimate an almost non-discriminatory empirical risk minimizer $\widehat{Y}$ by incorporating approximate non-discrimination constraints on the samples.
2. **Post-training correction**: With additional samples from $(\widehat{Y}, Y, A)$, derive a randomized predictor $\widetilde{Y}$ to further reduce discrimination.

4.1. Two step framework for binary predictors

We partition the training data consisting of n independent samples $S = \{(x_i, a_i, y_i) \sim \mathbb{P}(X, A, Y)\}_{i=1}^n$ into two subsets S_1 and S_2 to be used in Step 1 and Step 2, respectively[2]. For a predictor $h \in \mathcal{H}$ and $y, a \in \{0, 1\}$, recall notation for the population and sample group-conditional true and false positive rates $\gamma_{ya}(h)$ and $\gamma_{ya}^S(h)$ from (1) and (6), respectively. Additionally, for S_1 and S_2, let $n_{ya}^{S_k} = \sum_{i \in S_k} \mathbf{1}(y_i = y, a_i = a)$. In general the subsets *need not* be of equal size, but for simplicity, let $|S_1| = |S_2| = n/2$.

Step 1: Non-discrimination in training

For the first step, we estimate an empirical risk minimizing predictor $\widehat{Y} \in \mathcal{H}$, subject to the constraint that $\widehat{Y}$ be α_n-discriminatory on S_1, where α_n is a tunable hyperparameter:

$$\widehat{Y} = \operatorname*{argmin}_{h \in \mathcal{H}} \frac{2}{n} \sum_{i \in S_1} \ell^{01}(h(x_i), y_i) \tag{10}$$
$$\text{s.t.} \quad \Gamma^{S_1}(h) = \max_{y \in \{0,1\}} \left| \gamma_{y0}^{S_1}(h) - \gamma_{y1}^{S_1}(h) \right| < \alpha_n.$$

Step 2: Post-training correction

As Step 2, we propose a post hoc correction to $\widehat{Y}$ estimated in Step 1 for improved non-discrimination. Using samples in S_2, which are independent of S_1, we estimate the best randomized predictor $\widetilde{Y}$ derived from $(\widehat{Y}, A)$ (Definition 2). Let $\mathcal{P}(\widehat{Y})$ denote the set of randomized binary predictors that can be derived solely from $\mathbb{P}(\widehat{Y}, A, Y)$ and let $\widetilde{\alpha}_n$ be a tunable hyperparameter, then $\widetilde{Y}$ is given by,

$$\widetilde{Y} = \operatorname*{argmin}_{\widetilde{Y} \in \mathcal{P}(\widehat{Y})} \frac{2}{n} \sum_{i \in S_2} \mathbb{E}_{\widetilde{Y}} \ell^{01}(\widetilde{Y}(\widehat{y}_i, a_i), y_i) \tag{11}$$
$$\text{s.t.} \quad \Gamma^{S_2}(\widetilde{Y}) = \max_{y \in \{0,1\}} \left| \gamma_{y0}^{S_2}(\widetilde{Y}) - \gamma_{y1}^{S_2}(\widetilde{Y}) \right| < \widetilde{\alpha}_n,$$

where for a randomized predictor $\widetilde{Y}$, the group-conditional probabilities on a sample is defined to be $\gamma_{ya}^{S_2}(\widetilde{Y}) = \frac{1}{n_{ya}^{S_2}} \sum_{i \in S_2} \mathbb{E}_{\widetilde{Y}} \mathbf{1}(\widetilde{Y}_i = 1, Y_i = y, A_i = a)$. The above optimization problem is a finite sample adaptation of the post hoc correction in (3) proposed by Hardt et al. (2016).

---

2. We occasionally overload the S to also denote the indices $[n]$, e.g. $i \in S$ to denote the ith sample $(x_i, a_i, y_i) \in S$.

As with the post hoc correction on the population (3), estimating a predictor $\widetilde{Y} \in \mathcal{P}(\widehat{Y})$ derived from $(\widehat{Y}, A)$ is a simple optimization over the following four parameters that completely specify $\widetilde{Y}$,

$$\widetilde{p}_{\widehat{y}a} := \widetilde{p}_{\widehat{y}a}(\widetilde{Y}) = \mathbb{P}(\widetilde{Y} = 1 \mid \widehat{Y} = \widehat{y}, A = a) \text{ for } \widehat{y}, a \in \{0, 1\}. \tag{12}$$

In Section 4.2, we discuss how the post hoc correction step offers statistical advantages over the one-shot approach of using all of the training data for Step 1. Besides these statistical advantages, the post hoc correction step is also motivated from other practical considerations: (a) the derived predictors only need access to samples from $\mathbb{P}(\widehat{Y}, Y, A)$ and can be deployed without explicit access to predictive features X, and (b) the proposed correction step (11) can be easily optimized using ternary search and can be repeated multiple times as more and more samples from $\mathbb{P}(\widehat{Y}, Y, A)$ are seen by the system, without having to retrain the classifier from scratch.

4.2. Statistical guarantees

In this section, we discuss the statistical properties of the estimators $\widehat{Y}$ and $\widetilde{Y}$ from Step 1 and Step 2, respectively. We define the following notation for succinctly describing the quality of a predictor:

Definition 6 $\mathcal{Q}(L, \alpha) = \{h : \mathcal{X} \to \{0, 1\} : \Gamma(h) \leq \alpha, \mathcal{L}(h) \leq L\}$ *denotes the set of α-discriminatory binary predictors with loss L.*

The following theorem shows the statistical learnability of hypothesis classes $\mathcal{H}$ with respect to the best non-discriminatory predictor in $\mathcal{H}$ using the two-step framework. In the following results, for $y, a \in \{0, 1\}$, recall the notation $\mathrm{P}_{ya} = \mathbb{P}(Y = y, A = a)$ for the group-outcome probabilities. Additionally, for $a \in \{0, 1\}$, and let $VC(\mathcal{H})$ denote the Vapnik-Chervonenkis dimension of a hypothesis class $\mathcal{H}$.

Theorem 7 *Let $n = \Omega\left(\max_{ya} \frac{\log 1/\delta}{\mathrm{P}_{ya}}\right)$ and the hyperparameters satisfy $\alpha_n, \widetilde{\alpha}_n = \Theta\left(\max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathrm{P}_{ya}}}\right)$.*
For a binary hypothesis class $\mathcal{H}$, any distribution $\mathbb{P}(X, Y, A)$, and any $\delta \in (0, 1/2)$, if $Y^* \in \mathcal{H}$ is a non-discriminatory predictor, then with probability greater than $1 - \delta$, the output of the two step procedure $\widetilde{Y}$ satisfies the following for absolute constants C_1 and C_2,

$$\mathcal{L}(\widetilde{Y}) \leq \mathcal{L}(Y^*) + C_1 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathrm{P}_{ya}}}, \text{ and } \Gamma(\widetilde{Y}) \leq C_2 \max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathrm{P}_{ya}}}.$$

Thus, with $n = \Omega\left(\frac{VC(\mathcal{H})}{\epsilon^2} + \frac{1}{\alpha^2}\right) \frac{1}{\min_{ya} \mathrm{P}_{ya}}$ samples, and an appropriate choice of $\alpha_n, \widetilde{\alpha}_n$, the two step framework returns $\widetilde{Y} \in \mathcal{Q}(\mathcal{L}(Y^*) + \epsilon, \alpha)$ with high probability.

The proof is based on the following two Lemmas. The first is a statistical guarantee on the loss and non-discrimination after training in Step 1:

Lemma 8 *Under the conditions in Theorem 7, if $\alpha_n > 2 \max_{ya} \sqrt{\frac{\log 64/\delta}{n\mathrm{P}_{ya}}}$, then w.p. greater than $1 - \delta$, $\widehat{Y}$ from Step 1 satisfies*

$$\mathcal{L}(\widehat{Y}) \leq \mathcal{L}(Y^*) + C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}}, \text{ and } \Gamma(\widehat{Y}) \leq \alpha_n + C_2 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathrm{P}_{ya}}}.$$

8

The following second lemma ensures that if $\widehat{Y}$ from Step 1 is approximately non-discriminatory, the correction in the second step does not incur significant additional loss:

Lemma 9 *If h is an α-discriminatory binary predictor $h \in \mathcal{Q}(\mathcal{L}(h), \alpha)$, then the optimal 0-discriminatory derived predictor $\widetilde{Y}^*(h)$ from (3) using 0-1 loss satisfies $\widetilde{Y}^*(h) \in \mathcal{Q}(\mathcal{L}(h) + \alpha, 0)$.*

This lemma, along with the examples in Section 2 further motivates an integrated learning step, such as Step 1, where non-discrimination is explicitly encouraged in training.

Notice that, unlike to the guarantees for $\widetilde{Y}$ in Theorem 7, the upper bound on non-discrimination for $\widehat{Y}$ in Lemma 8 scales with the complexity of the hypothesis class $VC(\mathcal{H})$. In the Step 2, we search over a much restricted space of derived predictors $\mathcal{P}(\widehat{Y})$ which is essentially the convex hull of $|\mathcal{A}|$ conditional predictors of $\widehat{Y}$. This allows us to obtain a guarantee on non-discrimination for $\widetilde{Y}$ that does not scale with $VC(\mathcal{H})$. See Appendix C.3 for the remainder of the proof.

Finally, although Lemma 8 is only an upper bound on non-discrimination from Step 1, the following theorem shows that the level of non-discrimination using *just* the first step of the procedure can indeed grow with the complexity of the hypothesis class.

Theorem 10 *There exists a finite, binary hypothesis class $\mathcal{H}$ and a data distribution $\mathcal{D}$ such that with probability at least $1/2$, the classifier $\widehat{Y}$ learned from Step 1 using n samples from $\mathcal{D}$ is at least $\max_{y,a} \frac{3 \log \frac{|\mathcal{H}|-1}{5}}{4n \mathrm{P}_{ya}}$-discriminatory on the population.*

Theorem 10 suggests that the non-discrimination guarantee provided by Lemma 8 has the correct dependence on the problem parameters. Thus, for the intermediate predictor $\widehat{Y}$ from Step 1, without the post hoc correction, the tolerance for non-discrimination that can be guaranteed grows with the complexity of the hypothesis class (Lemma 8). On the other hand, when using the two step procedure, the sample complexity of ensuring that the final predictor $\widetilde{Y}$ is at most α-discriminatory is $\Omega(\max_{ya} 1/\alpha^2 \mathrm{P}_{ya})$ (Theorem 7) which *does not* depend on the complexity of the hypothesis class, and also matches the sample complexity of merely detecting α-discrimination from Lemma 4.

We also note that the sample complexity dependence on P_{ya} in Theorem 7 is unavoidable for our definition of approximate non-discrimination. If there is a rare group or group-outcome combination, we still need enough samples from that group to ensure that the loss and sample conditional distributions γ_{ya}^S generalize to the population for every $A = a, Y = y$. This is the same reason why the dependence on P_{ya} arises in Lemma 4. This bottleneck provides further incentive to actively seek samples and target labels for minority populations, which might otherwise be disregarded if non-discrimination were not a consideration.

5. Computational Intractability of Learning Non-discriminatory Predictors

The proposed procedure for learning non-discriminatory predictors from a finite sample is statistically optimal, but it is clearly computationally intractable for almost any interesting hypothesis class since the first step (10) involves minimizing the 0-1 loss. As is typically done with intractable learning problems, we therefore look to alternative loss functions and hypothesis classes in order to find a computationally feasible procedure.

A natural choice is the hypothesis class of real valued linear predictors with a convex loss function. In this case, we would like to have an efficient algorithm for finding a non-discriminatory predictor that has convex loss that is approximately as good as the loss of the best non-discriminatory

linear predictor. However, even in the case of binary A and Y and even with a convex loss, for real-valued predictors $h(x) \in \mathbb{R}$, the non-discrimination constraint in Defination 1 is extremely strong, requiring that the group-conditional true and false positive rates match at *every* threshold. In fact, the mere existence of a non-trivial (i.e. non-constant) linear predictor that is non-discriminatory requires a relatively special distribution. This is the case even when considering a real-valued analogue of α-approximate non-discrimination.

For binary targets $Y \in \{0, 1\}$, one could relax the problem one step further with a less restrictive non-discrimination requirement. Consider the class of linear predictors with a convex loss where only the sign of the predictor need be non-discriminatory. Unfortunately, using a result by Daniely (2015) even this is computationally intractable:

Theorem 11 *Let L^* be the hinge loss of the optimal linear predictor whose sign is non-discriminatory. Subject to the assumption that refuting random K-XOR formulas is computationally hard,[3] the learning problem of finding a possibly randomized function f such that $\mathcal{L}^{hinge}(f) \leq L^* + \epsilon$ such that sign(f) is α-discriminatory requires exponential time in the worst case for $\epsilon < \frac{1}{8}$ and $\alpha < \frac{1}{8}$.*

The proof goes through a reduction from the hardness of improper, agnostic PAC learning of HALF-SPACES. Given a distribution $\mathcal{D}$ over (X, Y) and the knowledge that there is a linear predictor which achieves 0-1 loss L^* on $\mathcal{D}$, we construct a new distribution $\widetilde{D}$ over $(\widetilde{X}, \widetilde{A}, \widetilde{Y})$ such that an approximately non-discriminatory predictor with small hinge loss can be used to make accurate predictions on $\mathcal{D}$, even if it is not a linear function. The distribution $\widetilde{\mathcal{D}}$ is identical to the original distribution $\mathcal{D}$ when conditioned on $\widetilde{A} = 1$, and is supported on only two points conditioned on $\widetilde{A} = 0$. The probabilities of the two points are constructed so that satisfying non-discrimination requires making accurate predictions on the $\widetilde{A} = 1$ population, and thus on $\mathcal{D}$. In particular, for parameters $\epsilon, \alpha < \frac{1}{8}$, the predictor will have 0-1 loss at most $\frac{15}{16}L^* + \frac{47}{128}$ on $\mathcal{D}$, which is bounded away from $\frac{1}{2}$ when $L^* < \frac{1}{10}$. Since Daniely (2015) prove that finding a predictor with accuracy bounded away from $\frac{1}{2}$ is hard in general, we conclude that the learning problem is computationally hard. See Appendix D for a complete proof.

To summarize, learning a non-discriminatory binary predictor with the 0-1 loss is hard; learning a real-valued linear predictor with respect to a convex loss function is problematic due to the potential inexistence of a non-trivial non-discriminatory linear predictor; and even requiring only that the sign of the linear predictor be non-discriminatory is computationally hard. A more significant relaxation of non-discrimination is therefore required to arrive at a computationally tractable learning problem.

6. Relaxing Non-discrimination for Tractable Estimation

Motivated by the hardness result in Section 5, we now proceed to relax the criterion of equalized odds. More precisely, we seek to identify a more tractable notion of non-discrimination based on *second order moments*. The most similar prior work to this is given by Zafar et al. (2016), who suggest a notion of non-discrimination based on *first order* moments. However, this may be viewed as a constraint on second order moments provided that A is binary. Namely, their notion amounts to relaxing the equalized odds constraint that $\mathbb{P}(\widehat{Y} = \widehat{y} \mid Y = y, A = 0) = \mathbb{P}(\widehat{Y} = \widehat{y} \mid Y = y, A = 1)$ to the constraint that $\mathbb{E}[\widehat{Y} \mid Y = y, A = 0] = \mathbb{E}[\widehat{Y} \mid Y = y, A = 1]$. They propose optimizing a convex loss subject to an approximation of this constraint. Their work is primarily applied and

---

3. See Daniely (2015) for a description of the problem.

gives no learning or non-discrimination guarantees for their learning rule, both of which we address in this section for a different relaxation.

In particular we propose the notion of *equalized correlations*. Equalized correlations is generally a weaker condition than equalized odds, but when considering the squared loss and when X, A, Y are jointly Gaussian, it is in fact equivalent.

Definition 12 (Equalized correlations) *We say that a real-valued predictor R satisfies* equalized correlations *or is* second-moment non-discriminatory *with respect to the protected attribute A and target Y, if R, A, and Y satisfy the following:*

$$\sigma_{RA}\sigma_Y^2 = \sigma_{RY}\sigma_{YA}, \tag{13}$$

where $\sigma_{\alpha\beta} = \mathbb{E}\left[(\alpha - \mathbb{E}[\alpha])(\beta - \mathbb{E}[\beta])\right]$ is the covariance between two random variables α and β.

Theorem 13 (Gaussian non-discrimination) *If X, A, and Y are jointly Gaussian, then the squared loss optimal (equalized odds) non-discriminatory predictor is linear. Furthermore, for all linear predictors in this setting, non-discrimination is equivalent to satisfying* (13).

Theorem 13 means in particular that in the Gaussian setting under the squared loss, the optimal predictor that is non-discriminatory in the sense of Definition 1 is the same as the optimal predictor that is non-discriminatory in the more relaxed sense of Definition 12. We stress that this is generally not the case for non-Gaussian scenarios and with losses other than the squared loss that may result in non-linear optimal predictors.

One may also consider intermediate notions of non-discrimination between equalized odds and equalized correlations. One such option is to require that the conditional covariance $\sigma_{RA|Y}$ vanishes. We do not elaborate further, except to note that it is immediate that equalized odds implies this to be the case and this in turn generally implies equalized correlations.

Since linear prediction can be thought of as the hallmark of Gaussian processes, one could therefore justify the relaxation of Definition 12 as being the appropriate notion of non-discrimination when we restrict our predictors to be linear. Linear predictors, especially under kernel transformations, are used in a wide array of applications. They thus form a practically relevant family of predictors where one would like to achieve non-discrimination. Therefore, in the remainder of this section, we develop a theory for non-discriminating linear predictors.

A particularly attractive feature of equalized correlations is that, for linear predictors, Equation (13) amounts to a linear constraint. With any convex loss $\ell(\cdot, \cdot)$, finding the optimal second-moment non-discriminatory predictor can be written as (where $\Sigma_{\cdot,\cdot}$ are covariance matrices):

$$\min_w \ \mathbb{E}\left[\ell\left(Y, w^T\left[\begin{smallmatrix} X \\ A \end{smallmatrix}\right]\right)\right] \qquad \text{s.t.} \quad w^T\left(\Sigma_{\left[\begin{smallmatrix} X \\ A \end{smallmatrix}\right],A}\sigma_Y^2 - \Sigma_{\left[\begin{smallmatrix} X \\ A \end{smallmatrix}\right],Y}\sigma_{YA}\right) = 0$$

This is a convex optimization problem with a single linear constraint, and is thus generally tractable.

In what follows, we take X to be a real-valued vector, A to be a scalar, and the target to be binary $Y \in \{\pm 1\}$, unless otherwise noted. We also commit to the squared loss $(R - Y)^2$ throughout. Without loss of generality, we assume that X, A, and Y all have zero mean. Recall the definition of the *optimal linear predictor*:

$$\widehat{R} = \underset{r(x,a)=w^T\left[\begin{smallmatrix} x \\ a \end{smallmatrix}\right]}{\arg\min} \ \mathbb{E}[(Y - r(X, A))^2] = \widehat{r}(X, A), \tag{14}$$

where one can determine that $\widehat{r}(x, a) = \widehat{w}^T \begin{bmatrix} x \\ a \end{bmatrix}$ with

$$\widehat{w} = \mathbb{E}\left[\begin{bmatrix} X \\ A \end{bmatrix}\begin{bmatrix} X^T & A^T \end{bmatrix}\right]^{-1}\mathbb{E}\left[\begin{bmatrix} X \\ A \end{bmatrix}Y\right] = \Sigma^{-1}_{\begin{bmatrix} X \\ A \end{bmatrix},\begin{bmatrix} X \\ A \end{bmatrix}}\Sigma_{\begin{bmatrix} X \\ A \end{bmatrix},Y}.$$

The optimal linear predictor may, in general, be discriminatory, so we define the *optimal second-moment non-discriminatory linear predictor* as follows:

$$R^\star = \underset{r(x,a)=w^T\begin{bmatrix} x \\ a \end{bmatrix}}{\arg\min}\mathbb{E}[(Y - r(X, A))^2] = r^\star(X, A) \qquad \text{s.t.}\quad \sigma_{r(X,A),A}\sigma^2_Y = \sigma_{r(X,A),Y}\sigma_{YA} \quad (15)$$

thus the predictor is constrained to be linear and to satisfy (13), which is a single linear constraint. We can give a closed-form expression for $R^\star$ (see Section E.2 in Appendix E for details). We have that $r^\star(x, a) = w^{\star T}\begin{bmatrix} x \\ a \end{bmatrix}$, where

$$w^\star = \Sigma^{-1}_{\begin{bmatrix} X \\ A \end{bmatrix},\begin{bmatrix} X \\ A \end{bmatrix}}\left(\Sigma_{\begin{bmatrix} X \\ A \end{bmatrix},Y} - \alpha \mathrm{v}\right)$$

where v is a vector encoding the non-discrimination constraint and α is a scalar defined as

$$\mathrm{v} = \Sigma_{\begin{bmatrix} X \\ A \end{bmatrix},A} - \Sigma_{\begin{bmatrix} X \\ A \end{bmatrix},Y}\sigma_{YA}/\sigma^2_Y \qquad \text{and} \qquad \alpha = \frac{\mathrm{v}^T\Sigma^{-1}_{\begin{bmatrix} X \\ A \end{bmatrix},\begin{bmatrix} X \\ A \end{bmatrix}}\Sigma_{\begin{bmatrix} X \\ A \end{bmatrix},Y}}{\mathrm{v}^T\Sigma^{-1}_{\begin{bmatrix} X \\ A \end{bmatrix},\begin{bmatrix} X \\ A \end{bmatrix}}\mathrm{v}}$$

Written as such, $R^\star$ is a function of X and A. Nevertheless, it turns out that this optimal non-discriminatory linear predictor can be derived, in the sense of Definition 2, from the optimal (possibly discriminatory) linear predictor $\widehat{R}$ of Equation 14 and without access to X individually.

Theorem 14 (Derived) *The second-moment non-discriminatory linear predictor minimizing the squared loss can be derived from the optimal least squares linear predictor $\widehat{R}$ and the joint (second moment) statistics of $(\widehat{R}, A, Y)$. Specifically,*

$$R^\star = \widehat{R} - \alpha\left(A - \widehat{R}\sigma_{YA}/\sigma^2_Y\right),$$

with

$$\alpha = \frac{\sigma_{YA} - \sigma_{\widehat{R},Y}\sigma_{YA}/\sigma^2_Y}{\sigma^2_A - 2(\sigma_{YA})^2/\sigma^2_Y + \sigma_{\widehat{R},Y}(\sigma_{YA})^2/(\sigma^2_Y)^2}.$$

Theorem 14 shows that, as far as the equalized correlation criterion is concerned, there is no penalty for first finding an optimal linear predictor and then correcting it. Consequently, this criterion easily enforceable on existing predictors. Intuitively, one must simply "subtract" any potential correlation one could derive about protected attributes from the prediction score and the outcome, and this can be entirely determined from the statistics of the optimal linear predictor $\widehat{R}$, the protected attribute A, and the outcome Y, without the need to know the extended set of attributes X. We emphasize that this result does not rely on any Gaussian assumptions, but simply on the fact that we have limited ourselves to linear predictors and the relaxed notion of non-discrimination. Finally, it is worth mentioning that a two-step procedure, as in Section 4, could be developed also for learning second-moment (approximately) non-discriminatory linear predictors from samples.

7. Conclusion

In this work we took the first steps toward a statistical and computational theory of learning non-discriminatory (equalized odds) predictors. We saw that post hoc correction might not be optimal and devised a statistically optimal two-step procedure, after observing that a straightforward ERM-type approach is not sufficient. Computationally, working with binary non-discrimination is essentially has hard as agnostically learning binary predictors, and so we should expect to have to resort to relaxations. We took the first step to this end in Section 6 where we considered a second moment relaxation of non-discrimination which leads to tractable learning. We hope this will not be the final word on learning non-discriminatory predictors and that this work will spur interest in further understanding our relaxation, suggesting other relaxations, and studying other computationally efficient procedures with provable guarantees.

References

Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

Amit Daniely. Complexity theoretic limitations on learning halfspaces. *arXiv preprint arXiv:1505.05800*, 2015.

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 441–448. ACM, 2014.

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

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016a.

Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016b.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

White House. Big data: A report on algorithmic systems, opportunity, and civil rights. 2016. URL https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *CoRR*, abs/1610.08452, 2016. URL http://dblp.uni-trier.de/db/journals/corr/corr1610.html#ZafarVGG16;http://arxiv.org/abs/1610.08452.

Appendix A. Deferred Proofs from Section 2

A.1. Proof of Example 1

We restate the example for convenience:

Example 1 *When the hypothesis class is unconstrained, for any $\epsilon \in (0, 1/4)$ there exists a distribution $\mathcal{D}_\epsilon$ such that **a)** the optimal non-discriminatory predictor Y^* with respect to the 0-1 loss has loss at most 2ϵ but **b)** for unrestricted Bayes optimal predictor $\widehat{Y}$ trained on 0-1 loss, the post hoc correction of $\widehat{Y}$ with 0-1 loss returns a predictor $\widetilde{Y}$ with loss at least 0.5.*

*A similar statement can also be made about predictors trained on hinge loss. For an unconstrained hypothesis class, for any $\epsilon \in (0, 1/4)$ and the same distribution $\mathcal{D}_\epsilon$, **a)** the optimal non-discriminatory predictor Y^* with respect to the hinge loss has loss at most 4ϵ but **b)** the post hoc correction of the Bayes optimal unrestricted predictor trained on hinge loss has loss 1.*

Consider the unconstrained hypothesis class of all (possibly randomized) functions from (X, A) to $\{0, 1\}$. Let $\mathcal{D}_\epsilon$ be the following distribution over (X, A, Y), with $X, A, Y \in \{0, 1\}$:

$$\mathbb{P}(Y = 1) = 0.5 \qquad \mathbb{P}(A = y \mid Y = y) = 1 - \epsilon \qquad \mathbb{P}(X = y \mid Y = y) = 1 - 2\epsilon \quad (16)$$

The graphical model representing this distribution is
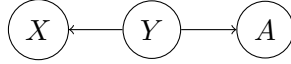
Clearly, $X \perp A \mid Y$, so $Y^* = X$ is non-discriminatory and achieves a 0-1 loss of 2ϵ. This same predictor achieves hinge loss 4ϵ. This predictor, being non-discriminatory, upper bounds the loss of the optimal non-discriminatory predictor with respect to the 0-1 and hinge losses.

The optimal predictor with respect to the 0-1 loss, which might be discriminatory, is in the convex hull (i.e. it might be randomized combination) of the sized mappings from $\{0, 1\} \times \{0, 1\} \mapsto \{0, 1\}$. The Bayes optimal predictor with respect to the 0-1 loss is the hypothesis

$$\widehat{h}(x, a) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \, \mathbb{P}(Y = y \mid X = x, A = a) \tag{17}$$

Given $a \in \{0, 1\}$, note that since $\epsilon < 1/4$

$$
\begin{aligned}
\mathbb{P}(Y = a \mid X = a, A = a) &= \frac{1}{1 + \frac{\mathbb{P}(A=a \mid Y=1-a)\mathbb{P}(X=a \mid Y=1-a)}{\mathbb{P}(A=a \mid Y=a)\mathbb{P}(X=a \mid Y=a)}} \\
&= \frac{1}{1 + \frac{(\epsilon)(2\epsilon)}{(1-\epsilon)(1-2\epsilon)}} > \frac{1}{2}.
\end{aligned}
\tag{18}
$$

Similarly

$$
\begin{aligned}
\mathbb{P}(Y = a \mid X = 1 - a, A = a) &= \frac{1}{1 + \frac{\mathbb{P}(A=a \mid Y=1-a)\mathbb{P}(X=1-a \mid Y=1-a)}{\mathbb{P}(A=a \mid Y=a)\mathbb{P}(X=1-a \mid Y=a)}} \\
&= \frac{1}{1 + \frac{\epsilon(1-2\epsilon)}{(1-\epsilon)(2\epsilon)}} > \frac{1}{2}.
\end{aligned}
\tag{19}
$$

Therefore, the Bayes optimal predictor is $\widehat{h}(X, A) = A$, which is 1-discriminatory, as

$$\mathbb{P}(\widehat{h}(X, A) = 1 \mid Y = y, A = 0) = 0 \qquad \text{but} \qquad \mathbb{P}(\widehat{h}(X, A) = 1 \mid Y = y, A = 1) = 1 \qquad (20)$$

Consider now the post-hoc correction $\widetilde{h}$ of $\widehat{h}$. The best non-discriminatory predictor $\widetilde{Y}$ derived from the joint distribution $(\widehat{h}, A, Y) \equiv (A, A, Y)$ is given by the following optimization problem

$$\widetilde{Y} = \underset{h}{\operatorname{argmin}} \, \mathcal{L}(h)$$

$$\text{s.t.} \ \gamma_{y0}(h) = \gamma_{y1}(h) \qquad \qquad \text{for } y = 0, 1$$

$$\begin{bmatrix} \gamma_{0a}(h) \\ \gamma_{1a}(h) \end{bmatrix} \in \operatorname{ConvHull}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \gamma_{0a}(\widehat{h}) \\ \gamma_{1a}(\widehat{h}) \end{bmatrix}, \begin{bmatrix} \gamma_{0a}(1 - \widehat{h}) \\ \gamma_{1a}(1 - \widehat{h}) \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) \qquad \text{for } a = 0, 1.$$

$$(21)$$

The first constraint requires that the resultant predictor be non-discriminatory. The second requires that the class conditional true positive and false positive rates of the predictor be in the convex hull of the constant 0 predictor, the constant 1 predictor, the predictor $\widehat{h}$ and its negative. This constraint is equivalent to requiring that $\widetilde{h}$ be derived from $\widehat{h}$. Because $\gamma_{y0}(\widehat{h}) = 0$ and $\gamma_{y1}(\widehat{h}) = 1$, the second constraint requires the predictor have equal true and false positive rates. As $\mathbb{P}(Y = 1) = 0.5$, 0.5 is optimal true and false positive rate and $\mathcal{L}(\widetilde{h}) = 0.5$.

Using the same distribution, but with $X, A, Y \in \{-1, 1\}$ instead of $\{0, 1\}$, the optimal non-discriminatory predictor with respect to the hinge loss is no worse than the predictor which returns X, achieving hinge loss 4ϵ. However, the Bayes optimal predictor with respect to the hinge loss is again that predictor which returns

$$\widehat{h}(x, a) = \underset{y \in \{0,1\}}{\operatorname{argmax}} \, \mathbb{P}(Y = y \mid X = x, A = a) = a.$$

By the same line of reasoning, the post hoc correction of $\widehat{h}$, which must have identical statistics for $A = 0$ and $A = 1$ is forced to have equal true and false positive rates, and thus the best derived predictor is identically 0 which has hinge loss 1.

A.2. Proof of Example 2

We restate the example for convenience:

Example 2 *Let $\mathcal{H}$ be the class of linear predictors with L^1 norm at most $\frac{1}{2} - 2\epsilon$, for some $\epsilon \in (2/25, 1/4)$. There exists a distribution $\mathcal{D}_\epsilon$ such that **a)** the optimal non-discriminatory predictor in $\mathcal{H}$ with respect to the squared loss has square loss at most $\frac{1}{16} + \frac{3\epsilon}{2} + 3\epsilon^2$, but **b)** the post hoc correction of the Bayes optimal square loss regressor in $\mathcal{H}$ returns a constant predictor which has (trivial) square loss of $1/4$.*

*Similarly, for the class $\mathcal{H}$ of sparse linear predictors, for any $\epsilon \in (0, 1/4)$, there exists a distribution $\mathcal{D}_\epsilon$ such that **a)** the optimal non-discriminatory predictor in $\mathcal{H}$ with respect to the squared loss has square loss at most $2\epsilon - 4\epsilon^2$, but **b)** the post hoc correction of the Bayes optimal squared loss regressor in $\mathcal{H}$ again returns a constant predictor which has (trivial) square loss of $1/4$.*

In this example, we consider the squared loss and the hypothesis class of linear predictors with L^1 norm at most $\frac{1}{2} - 2\epsilon$ for some $\epsilon \in (2/25, 1/4)$:

$$\mathcal{H} = \left\{ w_1 X + w_2 A + b : |w_1| + |w_2| \le \frac{1}{2} - 2\epsilon \right\}.$$

With this parameter ϵ, we use the same distribution as in the proof of Theorem 1:

$$\mathbb{P}(Y = 1) = 0.5 \qquad \mathbb{P}(A = y \mid Y = y) = 1 - \epsilon \qquad \mathbb{P}(X = y \mid Y = y) = 1 - 2\epsilon$$

Since $X \perp\!\!\!\perp A \mid Y$, any linear function of X only will be non-discriminatory and $h(X) = \left(\frac{1}{2} - 2\epsilon\right) X + \frac{1}{4} + \epsilon$ has the required L^1 norm and achieves squared loss $\frac{1}{16} + \frac{3\epsilon}{2} + 3\epsilon^2$. This predictor, being non-discriminatory, upper bounds the squared loss of the optimal non-discriminatory predictor.

To derive the optimal potentially discriminatory predictor, it will be useful to begin by calculating the covariances between each of the variables:

$$\mathbb{E}[X] = \mathbb{E}[A] = \mathbb{E}[Y] = \frac{1}{2} \tag{22}$$

$$\mathbb{E}[X^2] = \mathbb{E}[A^2] = \mathbb{E}[Y^2] = \frac{1}{2} \tag{23}$$

$$\mathbb{E}[XA] = \frac{1}{2} - \frac{3}{2}\epsilon + 2\epsilon^2 \tag{24}$$

$$\mathbb{E}[XY] = \frac{1 - 2\epsilon}{2} \tag{25}$$

$$\mathbb{E}[AY] = \frac{1 - \epsilon}{2} \tag{26}$$

The optimal predictor optimizes

$$\widehat{h} = \underset{w_1, w_2, b}{\operatorname{argmin}} \, \mathbb{E}\left[(w_1 X + w_2 A + b - y)^2\right]$$

$$\text{s.t.} \quad |w_1| + |w_2| \leq \frac{1}{2} - 2\epsilon. \tag{27}$$

Forming the Lagrangian:

$$\mathcal{L}(w_1, w_2, b, \lambda) = \mathbb{E}\left[(w_1 X + w_2 A + b - y)^2\right] + \lambda\left(|w_1| + |w_2| - \frac{1}{2} - 2\epsilon\right)$$

$$= w_1^2 \mathbb{E}[X^2] + w_2^2 \mathbb{E}[A^2] + b^2 + \mathbb{E}[Y^2] + 2w_1 w_2 \mathbb{E}[XA] + 2w_1 b \mathbb{E}[X]$$

$$- 2w_1 \mathbb{E}[XY] + 2w_2 b \mathbb{E}[A] - 2w_2 \mathbb{E}[AY] - 2b \mathbb{E}[Y] + \lambda\left(|w_1| + |w_2| - \frac{1}{2} - 2\epsilon\right)$$

$$= \frac{w_1^2 + w_2^2 + 1}{2} + b^2 + w_1 w_2(1 - 3\epsilon + 4\epsilon^2) + w_1 b - w_1(1 - 2\epsilon) + w_2 b$$

$$- w_2(1 - \epsilon) - b + \lambda\left(|w_1| + |w_2| - \frac{1}{2} - 2\epsilon\right). \tag{28}$$

At the following values:

$$w_1 = 0 \qquad w_2 = \frac{1}{2} - 2\epsilon \qquad b = \frac{1}{4} + \epsilon \qquad \lambda = \frac{1}{4} \tag{29}$$

the subdifferential of $\mathcal{L}$ contains 0 for any $\epsilon \in (2/25, 1/4)$. Looking term by term we have:

$$\frac{\partial \ell}{\partial w_1} = w_1 + w_2(1 - 3\epsilon + 4\epsilon^2) + b - (1 - 2\epsilon) + \lambda \operatorname{sign}(w_1) \tag{30}$$

$$= \left(\frac{1}{2} - 2\epsilon\right)(1 - 3\epsilon + 4\epsilon^2) + \frac{1}{4} + \epsilon - (1 - 2\epsilon) + \frac{1}{4}\operatorname{sign}(0) \tag{31}$$

$$= \frac{1}{4}\operatorname{sign}(0) - 8\epsilon^3 + 8\epsilon^2 - \frac{\epsilon}{2} - \frac{1}{4}, \tag{32}$$

where $\text{sign}(0)$ is an arbitrary value in $[-1, 1]$ which is the subdifferential of $f(z) = |z|$ at 0. For any $\epsilon \in (2/25, 1/4)$

$$\left| -8\epsilon^3 + 8\epsilon^2 - \frac{\epsilon}{2} - \frac{1}{4} \right| < \frac{1}{4} \implies 0 \in \frac{\partial \ell}{\partial w_1} \tag{33}$$

Furthermore,

$$\begin{aligned} \frac{\partial \ell}{\partial w_2} &= w_1(1 - 3\epsilon + 4\epsilon^2) + w_2 + b - (1 - \epsilon) + \lambda \text{sign}(w_2) \\ &= \frac{1}{2} - 2\epsilon + \frac{1}{4} + \epsilon - (1 - \epsilon) + \frac{1}{4} = 0, \end{aligned} \tag{34}$$

and

$$\begin{aligned} \frac{\partial \ell}{\partial b} &= 2b + w_1 + w_2 - 1 \\ &= \frac{1}{2} + 2\epsilon + \frac{1}{2} - 2\epsilon - 1 = 0. \end{aligned} \tag{35}$$

This proves that $\widehat{h}(X, A) = \left(\frac{1}{2} - 2\epsilon\right) A + \frac{1}{4} + \epsilon$ is the Bayes optimal predictor in $\mathcal{H}$ with respect to the squared loss. Furthermore, the random variable is supported on only two points: $\frac{1}{4} + \epsilon$ when $A = 0$ and $\frac{3}{4} - \epsilon$ when $A = 1$. It is clear that $\widehat{h}$ is not independent of A conditioned on Y.

Since $\widehat{h}$ is a deterministic function of A, the post hoc correction $\widetilde{h}$, which must be independent of A conditioned on Y, is forced to be independent of A, and consequently Y. Thus, $\widetilde{h} \equiv 0.5$ is the best possible derived predictor, achieving square loss $1/4$.

Considering the class of 1-sparse linear predictors, the predictor $h(X, A) = (1 - 4\epsilon)X + 2\epsilon$, being conditionally independent of A is non-discriminatory and achieves squared loss $2\epsilon - 4\epsilon^2$, upper bounding the loss of the optimal non-discriminatory 1-sparse linear predictor. Without regard for non-discrimination, the optimal hypothesis in the class with respect to the squared loss is $\widehat{h}(X, A) = (1 - 2\epsilon)A + \epsilon$. This post hoc correction of this predictor suffers from the same issue as in the bounded norm case, resulting in a predictor that can have squared loss no better than $1/4$.

Appendix B. Deferred Proofs From Section 3

Recall the notation $\Gamma_{ya}(h) = \max_y |\gamma_{y0}(h) - \gamma_{y1}(h)|$ and $\Gamma_{ya}^S(h) = \max_y |\gamma_{y0}^S(h) - \gamma_{y1}^S(h)|$ from Definition 3. To avoid clutter, we sometimes drop the dependence on h for γ_{ya} when h is evident from the context. Also, recall that $S = \{(x_i, y_i, a_i) : i \in [n]\} \sim \mathbb{P}^n(X, Y, A)$ and $P_{ya} = \mathbb{P}(Y = y, A = a)$.

B.1. Proof of Lemma 4

Lemma 4 *Given n i.i.d. samples S, $\forall \alpha \in (0, 1), \delta \in (0, 1/2)$, if $n > \frac{16 \log 32/\delta}{\alpha^2 \min_{ya} P_{ya}}$, then with probability greater than $1 - \delta$, T satisfies,*

$$T\left(\widehat{Y}, S, \frac{\alpha}{2}\right) = \begin{cases} 0 & \text{if } \widehat{Y} \text{ is 0-discriminatory on population} \\ 1 & \text{if } \widehat{Y} \text{ is at least } \alpha\text{-discriminatory on population.} \end{cases}$$

Proof Recall that $T(\widehat{Y}, S, \alpha) = \mathbf{1}\left(\Gamma^S(\widehat{Y}) > \alpha\right)$. Let $\alpha_n > 0$ be any parameter chosen to satisfy,

$$2 \max_{ya} \sqrt{\frac{\log 32/\delta}{n \mathrm{P}_{ya}}} < \alpha_n < \alpha - 2 \max_{ya} \sqrt{\frac{\log 32/\delta}{n \mathrm{P}_{ya}}}. \tag{36}$$

Then the following results readily follow from Lemma 5,

1. If $\widehat{Y}$ is non-discriminatory, i.e. $\Gamma(\widehat{Y}) = 0$, then

$$\mathbb{P}\left(T(\widehat{Y}, S, \alpha_n) = 1\right) = \mathbb{P}(\Gamma^S(\widehat{Y}) > \alpha_n) \leq \mathbb{P}\left(\Gamma^S(\widehat{Y}) > \Gamma(\widehat{Y}) + 2 \max_{ya} \sqrt{\frac{\log 32/\delta}{n \mathrm{P}_{ya}}}\right) \leq \frac{\delta}{2}.$$

2. Similarly, suppose is $\widehat{Y}$ is at least α-discriminatory on the population, i.e. $\Gamma(\widehat{Y}) \geq \alpha$, then

$$\mathbb{P}\left(T(\widehat{Y}, S, \alpha_n) = 0\right) = \mathbb{P}\left(\Gamma^S(\widehat{Y}) \leq \alpha_n\right) \leq \mathbb{P}\left(\Gamma^S(\widehat{Y}) \leq \Gamma(\widehat{Y}) - 2 \max_{ya} \sqrt{\frac{\log 32/\delta}{n \mathrm{P}_{ya}}}\right) \leq \frac{\delta}{2}.$$

Thus, if $\frac{\alpha}{2} > 2 \max_{ya} \sqrt{\frac{\log 32/\delta}{n \mathrm{P}_{ya}}}$, then $\alpha_n = \frac{\alpha}{2}$ satisfies (36) and Lemma 4 follows for $T(\widehat{Y}, S, \frac{\alpha}{2})$. ∎

B.2. Proof of Lemma 5

Lemma 5 *For $\delta \in (0, 1/2)$ and a binary predictor h, if $n > \frac{8 \log 8/\delta}{\min_{ya} \mathrm{P}_{ya}}$, then*

$$\mathbb{P}\left(\left|\Gamma(h) - \Gamma^S(h)\right| > 2 \max_{ya} \sqrt{\frac{\log 16/\delta}{n \mathrm{P}_{ya}}}\right) \leq \delta. \tag{37}$$

Proof Recall that $n_{ya}^S = \sum_i \mathbf{1}(y_i = y, a_i = a)$. With slight abuse of notation, we define random variables $S_{ya} = \{i : y_i = y, a_i = a\}$.

We then have $\gamma_{ya}^S(h) | S_{ya} = \frac{\sum_{j \in S_{ya}} h(x_j, a_j)}{n_{ya}^S} \sim \frac{1}{n_{ya}^S} \text{Binomial}(\gamma_{ya}, n_{ya}^S)$ with $\mathbb{E}[\gamma_{ya}^S | S_{ya}] = \gamma_{ya}$.

$$\mathbb{P}\left(|\gamma_{ya}^S - \gamma_{ya}| > t\right) \overset{(a)}{=} \sum_{S_{ya}} \mathbb{P}\left(|\gamma_{ya}^S - \gamma_{ya}| > t | S_{ya}\right) \mathbb{P}\left(S_{ya}\right)$$

$$\leq \mathbb{P}\left(n_{ya}^S < \frac{n \mathrm{P}_{ya}}{2}\right) + \sum_{S_{ya}: n_{ya}^S \geq \frac{n \mathrm{P}_{ya}}{2}} \mathbb{P}\left(|\gamma_{ya}^S - \gamma_{ya}| > t | S_{ya}\right) \mathbb{P}\left(S_{ya}\right)$$

$$\overset{(b)}{\leq} \exp\left(-\frac{n \mathrm{P}_{ya}}{8}\right) + \sum_{S_{ya}: n_{ya}^S \geq \frac{n \mathrm{P}_{ya}}{2}} 2 \exp\left(-2 t^2 n_{ya}^S\right) \mathbb{P}(S_{ya})$$

$$\overset{(c)}{\leq} \frac{\delta}{8} + 2 \exp\left(-t^2 n \mathrm{P}_{ya}\right), \tag{38}$$

where in (a) the summation is over all 2^n possible configurations of $S_{ya} \subset [n]$, (b) follow from Chernoff bound on $n_{ya}^S \sim \text{Binomial}(n, P_{ya})$ and Hoeffding's bound on $\gamma_{ya}^S | S_{ya} \sim \frac{1}{n_{ya}^S}\text{Binomial}(n_{ya}^S, \gamma_{ya})$, and (c) follows from the condition on nP_{ya} in Lemma 5.

Further, for $y \in \{0, 1\}$, using a series of triangle inequality,

$$\left||\gamma_{y0}^S - \gamma_{y1}^S| - |\gamma_{y0} - \gamma_{y1}|\right| \le |\gamma_{y0}^S - \gamma_{y1}^S - \gamma_{y0} + \gamma_{y1}| \le |\gamma_{y0}^S - \gamma_{y0}^S| + |\gamma_{y1}^S - \gamma_{y1}|, \text{ and}$$

$$\begin{aligned}
\mathbb{P}_S\left(\left||\gamma_{y0}^S - \gamma_{y1}^S| - |\gamma_{y0} - \gamma_{y1}|\right| > 2t\right) &\le \mathbb{P}_S\left(|\gamma_{y0}^S - \gamma_{y0}| + |\gamma_{y1}^S - \gamma_{y1}| > 2t\right) \\
&\overset{(a)}{\le} \mathbb{P}_S\left(|\gamma_{y0}^S - \gamma_{y0}| > t\right) + \mathbb{P}_S\left(|\gamma_{y1}^S - \gamma_{y1}| > t\right) \\
&\le \frac{\delta}{4} + 4\exp\left(-t^2 nP_{ya}\right) \overset{(b)}{\le} \frac{\delta}{2},
\end{aligned} \tag{39}$$

where (a) follows from union bound, and (b) follows from (38) using $t = \max_a \sqrt{\frac{\log 16/\delta}{nP_{ya}}}$. The lemma follows from collecting the failure probabilities for $y = 0, 1$. ∎

Appendix C. Deferred Proofs From Section 4

We use the notation $A \le_\delta B$ to denote that $A \le B$ holds with probability greater than $1 - \delta$. Recall the notation $\Gamma_{ya}(h) = \max_y |\gamma_{y0}(h) - \gamma_{y1}(h)|$ and $\Gamma_{ya}^S(h) = \max_y |\gamma_{y0}^S(h) - \gamma_{y1}^S(h)|$ from Definition 3. To avoid clutter, we sometimes drop the dependence on h for γ_{ya} when h is evident from the context. Finally, in this section C, C_1 and C_2 denote absolute constants that are not necessarily the same at each occurrence.

C.1. Proof of Lemma 8

The following intermediate lemma is used in proof of Lemma 8.

Lemma 15 *Let $\mathcal{H}_{\alpha_n}^{S_1} = \{h \in \mathcal{H} : \Gamma^{S_1}(h) \le \alpha_n\}$ denote the subset of hypothesis that satisfy the constraints in (10). If $\forall(y, a), nP_{ya} > 16\log 8/\delta$ and α_n in (10) satisfies $\alpha_n \ge 2\max_{ya}\sqrt{\frac{2\log 64/\delta}{nP_{ya}}}$, then with probability greater than $1 - \frac{\delta}{4}$, $Y^* \in \mathcal{H}_{\alpha_n}^{S_1}$ for all $Y^* \in \mathcal{Q}(\mathcal{L}(Y^*), 0) \cap \mathcal{H}$.*

Proof Given $Y^* \in \mathcal{H} \cap \mathcal{Q}(L^*, 0)$,

$$\begin{aligned}
\mathbb{P}(Y^* \notin \mathcal{H}_{\alpha_n}^{S_1}) = \mathbb{P}(\Gamma^{S_1}(Y^*) > \alpha_n) &\overset{(a)}{\le} \mathbb{P}\left(\Gamma^{S_1}(Y^*) > 2\max_{ya}\sqrt{\frac{2\log 64/\delta}{nP_{ya}}}\right) \\
&\overset{(b)}{\le} \mathbb{P}\left(\Gamma^{S_1}(Y^*) > \Gamma(Y^*) + 2\max_{ya}\sqrt{\frac{2\log 64/\delta}{nP_{ya}}}\right) \\
&\overset{(c)}{\le} \frac{\delta}{4},
\end{aligned} \tag{40}$$
$$\tag{41}$$

where (a) follows from the condition on α_n, (b) follows from $\Gamma(Y^*) = 0$ assumption, and (c) follows from Lemma 5 as $|S_1| = n/2$. ∎

Lemma 8 *Under the conditions in Theorem 7, if $\alpha_n > 2 \max_{ya} \sqrt{\frac{2 \log 64/\delta}{n P_{ya}}}$, then w.p. greater than $1 - \delta$, $\widehat{Y}$ from Step 1 satisfies*

$$\mathcal{L}(\widehat{Y}) \leq \mathcal{L}(Y^*) + C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}}, \text{ and } \Gamma(\widehat{Y}) \leq \alpha_n + C_2 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n P_{ya}}}.$$

Proof Recall that the training data for Step 1 are denoted by $S_1 = \{(x_i, a_i, y_i) : i \in [n/2]\} \sim \mathbb{P}^{n/2}(X, A, Y)$ and $P_{ya} = \mathbb{P}(Y = y, A = a)$. From Using Hoeffding's inequality on the empirical 0-1 loss $\mathcal{L}^{S_1}(h)$, and using concentration results for $\Gamma^{S_1}(h)$ from Lemma 5, respectively, the following holds for $\delta \in (0, 1/2)$ and $\min_{ya} n P_{ya} > 16 \log 8/\delta$.

$$\left| \mathcal{L}(h) - \mathcal{L}^{S_1}(h) \right| \leq_{\delta/4} \sqrt{\frac{\log 8/\delta}{n}}, \text{ and } \left| \Gamma(h) - \Gamma^{S_1}(h) \right| \leq_{\delta/4} 2 \max_{ya} \sqrt{\frac{2 \log 64/\delta}{n P_{ya}}}. \quad (42)$$

Using (42) and the standard VC dimension uniform bound (Bousquet et al., 2004), the following holds with high probability for absolute constants C_1 and C_2,

$$\begin{aligned}
\left| \mathcal{L}(\widehat{Y}) - \mathcal{L}^{S_1}(\widehat{Y}) \right| &\leq_{\delta/4} C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}}, \quad \text{and} \\
\left| \Gamma(\widehat{Y}) - \Gamma^{S_1}(\widehat{Y}) \right| &\leq_{\delta/4} C_2 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n P_{ya}}}.
\end{aligned} \quad (43)$$

Finally, from Lemma 15, with probability greater than $1 - \delta/4$, any 0-discriminatory $Y^* \in \mathcal{H}$ is in the feasible set for Step 1 in (10), and thus from the optimality of $\widehat{Y}$, $\mathcal{L}^{S_1}(\widehat{Y}) \leq_{\delta/4} \mathcal{L}^{S_1}(Y^*) \leq_{\delta/4} \mathcal{L}(Y^*) + C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}}$. Thus,

$$\mathcal{L}(\widehat{Y}) \leq_{\delta/4} \mathcal{L}^{S_1}(\widehat{Y}) + C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}} \leq_{\delta/2} \mathcal{L}(Y^*) + 2C_1 \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n}},$$

$$\Gamma(\widehat{Y}) \leq_{\delta/4} \alpha_n + C_2 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n P_{ya}}}. \quad (44)$$

The lemma follows from combining the failure probabilities in the above equation. ■

C.2. Proof of Lemma 9

Lemma 9 *If h is an α-discriminatory binary predictor $h \in \mathcal{Q}(\mathcal{L}(h), \alpha)$, then the optimal 0-discriminatory derived predictor $\widetilde{Y}^*(h)$ from (3) using 0-1 loss satisfies $\widetilde{Y}^*(h) \in \mathcal{Q}(\mathcal{L}(h) + \alpha, 0)$.*
Proof The intuition is to conservatively bound the true and false positive rates of the non-discriminatory derived predictor using the class conditional rates for h. In the case of binary predictors, $\widetilde{Y}$ being derived from h is equivalent to requiring that

$$\left(\gamma_{0a}(\widetilde{Y}(h)), \gamma_{1a}(\widetilde{Y}(h)) \right) \in \text{Conv} \left((0, 0), (1, 1), (\gamma_{0a}(h), \gamma_{1a}(h)), (1 - \gamma_{0a}(h), 1 - \gamma_{1a}(h)) \right) \quad (45)$$

In the figure below, $\widetilde{Y}^*(h)$ is the actual optimal derived non-discriminatory predictor, but we estimate it conservatively using the worse of the class conditional true and false positive rates of h:
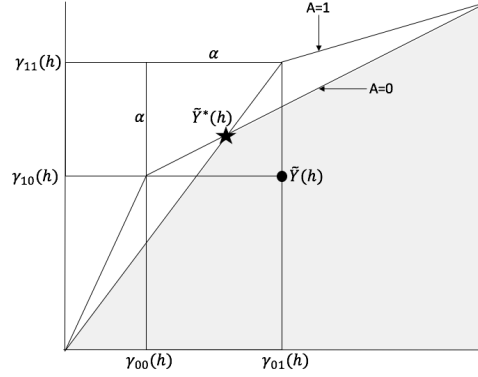
Without loss of generality, assume $\gamma_{1a}(h) \geq 0.5$ and $\gamma_{0a}(h) \leq 0.5$ for all a (the hypothesis is at least as good as chance). Consider the predictor $\widetilde{Y}$ such that for both $a \in \{0,1\}$,

$$\left(\gamma_{0a}(\widetilde{Y}), \gamma_{1a}(\widetilde{Y})\right) = (\max(\gamma_{00}, \gamma_{01}), \min(\gamma_{10}, \gamma_{11})) \in \mathrm{Conv}\left((0,0), (1,1), (\gamma_{0a}(h), \gamma_{1a}(h))\right)$$

that is, $\widetilde{Y}$ has the greater of the two false positive rates and lesser of the two true positive rates for both classes $A = 1$ and $A = 0$. Additionally, Lemma 9 requires that $\forall y, |\gamma_{y1}(h) - \gamma_{y0}(h)| \leq \alpha$. Thus, for $a \in \{0,1\}$,

$$
\begin{aligned}
\gamma_{0a}(\widetilde{Y}) - \gamma_{0a}(h) &= \max_{a'} \gamma_{0a'}(h) - \gamma_{0a}(h) \leq \alpha, \text{ and} \\
\gamma_{1a}(h) - \gamma_{1a}(\widetilde{Y}) &= \gamma_{1a}(h) - \min_{a'} \gamma_{1a'}(h) \leq \alpha.
\end{aligned}
\tag{46}
$$

Clearly, this choice of $\widetilde{Y}$ is both non-discriminatory (as $\gamma_{ya}(\widetilde{Y})$ is set independent of a for all y), as well as derived (as $\gamma_{ya}(\widetilde{Y})$ satisfy (45)). Thus $\widetilde{Y}$ is a feasible point for (3), and we have

$$
\begin{aligned}
\mathbb{E}[\ell^{01}(\widetilde{Y}^*(h))] &\leq \mathbb{E}[\ell^{01}(\widetilde{Y})] = \sum_{a \in \{0,1\}} \mathrm{P}_{0a} \gamma_{0a}(\widetilde{Y}) + \sum_{a \in \{0,1\}} \mathrm{P}_{1a}(1 - \gamma_{1a}(\widetilde{Y})) \\
&\overset{(a)}{\leq} \sum_{a \in \{0,1\}} \mathrm{P}_{0a} \left(\gamma_{0a}(h) + \alpha\right) + \sum_{a \in \{0,1\}} \mathrm{P}_{1a} \left(1 - (\gamma_{1a}(h) - \alpha)\right) \\
&\leq \mathbb{E}[\ell^{01}(h)] + \alpha,
\end{aligned}
\tag{47}
$$

where (a) follows from (47). ∎

C.3. Proof of Theorem 7

The following supporting lemma on concentration of non-discrimination for randomized predictors is used in the proof of Theorem 7.

Lemma 16 *For $\delta \in (0, 1/2)$ and $h \in \mathcal{H}$, if $n\mathrm{P}_{ya} > 16 \log 8/\delta$, For any randomized predictor $\widetilde{Y}$ derived from $(\widehat{Y}, A)$, i.e. $\widetilde{Y} \in \mathcal{P}(\widehat{Y})$, satisfies the following for $|S_2| = n/2$ iid sampels:*

$$\left|\mathcal{L}(\widetilde{Y}) - \mathcal{L}^{S_2}(\widetilde{Y})\right| \leq_\delta \sqrt{\frac{\log 2/\delta}{n}}, \text{ and } |\Gamma(\widetilde{Y}) - \Gamma^{S_2}(\widetilde{Y})| \leq_\delta 2 \max_{ya} \sqrt{\frac{2 \log 16/\delta}{n\mathrm{P}_{ya}}}. \quad (48)$$

Proof The proof essentially follows the same arguments that were used for (42) and Lemma 5.

For a randomized predictor $\widetilde{Y}$,

$$\mathcal{L}^{S_2}(\widetilde{Y}) - \mathcal{L}(\widetilde{Y}) = \frac{2}{n} \sum_{i \in S_2} \mathbb{E}_{\widetilde{Y}} \ell(\widetilde{Y}(\widehat{y}_i, a_i), y_i) - \mathbb{E}_{X,A,Y} \mathbb{E}_{\widetilde{Y}} \ell(\widetilde{Y}, Y)$$

Here $\mathcal{L}^{S_2}(\widetilde{Y}) - \mathcal{L}(\widetilde{Y})$ is merely a sum of $n/2$ independent and $[0, 1]$ bounded random variables and Hoeffdings bound can be applied to get the required concentration on $\mathcal{L}^{S_2}$.

Similarly, for any randomized predictor $\widetilde{Y}$, the conditional random variable $\gamma_{ya}^{S_2}(\widetilde{Y})|S_{2ya} = \frac{\sum_{j \in S_{2ya}} \mathbb{E}_{\widetilde{Y}} \widetilde{Y}(\widehat{y}_j, a_j)}{n_{ya}^{S_2}}$ is a sum of $[0, 1]$ bounded random variables with mean $\mathbb{E}[\gamma_{ya}^{S_2}(\widetilde{Y})|S_{2ya}] = \gamma_{ya}(\widetilde{Y})$, the proof of Lemma 5 can be repeated verbatim for the randomized prediction where instead of the Hoeffdings' bound on Binomial random variables, we use the identical Hoeffdings' bound for $[0, 1]$ bounded random variables. ∎

Theorem 7 *Let $n = \Omega\left(\max_{ya} \frac{\log 1/\delta}{\mathrm{P}_{ya}}\right)$ and the hyperparameters satisfy $\alpha_n, \widetilde{\alpha}_n = \Theta\left(\max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathrm{P}_{ya}}}\right)$.*

For a binary hypothesis class $\mathcal{H}$, any distribution $\mathbb{P}(X, Y, A)$, and any $\delta \in (0, 1/2)$, if $Y^* \in \mathcal{H}$ is a non-discriminatory predictor, then with probability greater than $1 - \delta$, the output of the two step procedure $\widetilde{Y}$ satisfies the following for absolute constants C_1 and C_2,

$$\mathcal{L}(\widetilde{Y}) \leq \mathcal{L}(Y^*) + C_1 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathrm{P}_{ya}}}, \text{ and } \Gamma(\widetilde{Y}) \leq C_2 \max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathrm{P}_{ya}}}.$$

Thus, with $n = \Omega\left(\frac{VC(\mathcal{H})}{\epsilon^2} + \frac{1}{\alpha^2}\right) \frac{1}{\min_{ya} \mathrm{P}_{ya}}$ samples, and an appropriate choice of $\alpha_n, \widetilde{\alpha}_n$, the two step framework returns $\widetilde{Y} \in \mathcal{Q}(\mathcal{L}(Y^*) + \epsilon, \alpha)$ with high probability.

Proof We begin with the following result from from Lemma 16 which shows the concentration of loss and discrimination in radomized derived predictors: if $n\mathrm{P}_{ya} > 16 \log 8/\delta$, for any randomized predictor $\widetilde{h}$ derived from $(\widehat{Y}, A)$, i.e. $\widetilde{h} \in \mathcal{P}(\widehat{Y})$, satisfies the following:

$$\left|\mathcal{L}(\widetilde{h}) - \mathcal{L}^{S_2}(\widetilde{h})\right| \leq_{\delta/4} \sqrt{\frac{\log 8/\delta}{n}}, \text{ and } |\Gamma(\widetilde{h}) - \Gamma^{S_2}(\widetilde{h})| \leq_{\delta/4} 2 \max_{ya} \sqrt{\frac{2 \log 64/\delta}{n\mathrm{P}_{ya}}}. \quad (49)$$

VC dimension of $\mathcal{P}(\widehat{Y})$: If $|\widehat{\mathcal{Y}}|$ and $|\mathcal{A}|$ are finite, consider a finite hypothesis class denoted by $\mathcal{H}_{\widehat{Y},A}$, that includes all deterministic mappings from $(\widehat{Y}, A)$ to binary values $\{0, 1\}$. If $\widehat{Y}$ and A are both binary, then there are $4^2 = 16$ such mappings $\mathcal{H}_{\widehat{Y},A} = \{\widetilde{h} : \{0, 1\} \times \{0, 1\} \to \{0, 1\}\}$.

Further, recall that the feasible set of (randomized) binary predictors derived from $\mathbb{P}(\widehat{Y}, Y, A)$ is denoted by $\mathcal{P}(\widehat{Y})$, and any $\widetilde{Y} \in \mathcal{P}(\widehat{Y})$ is completely specified by four parameters, $\{\widetilde{p}_{\widehat{y},a}(\widetilde{Y}) = \mathbb{P}(\widetilde{Y} = 1|\widehat{Y} = \widehat{y}, A = a) : \widehat{y}, a \in \{0, 1\}\}$. With the above definition of $\mathcal{H}_{\widehat{Y},A}$, any such randomized derived predictor $\widetilde{Y} \in \mathcal{P}(\widehat{Y})$ derived from $(\widehat{Y}, A)$ is in the convex hull of $\mathcal{H}_{\widehat{Y},A}$. This implies, $VC(\mathcal{P}(\widehat{Y})) = VC(\text{conv}(\mathcal{H}_{\widehat{Y},A})) = \log 16$ which is a constant.

Thus, for $\forall \widetilde{Y} \in \mathcal{P}(\widehat{Y})$ estimated from Step 2 in (11), using the standard VC dimension uniform bound over $\mathcal{P}(\widehat{Y})$ (Bousquet et al., 2004) along with (49) we have the following,

$$\mathcal{L}(\widetilde{Y}) \leq_{\delta/4} \mathcal{L}^{S_2}(\widetilde{Y}) + C_1 \sqrt{\frac{\log 1/\delta}{n}}, \text{ and } \Gamma(\widetilde{Y}) \leq_{\delta/4} \widetilde{\alpha}_n + C_2 \max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathbf{P}_{ya}}}. \tag{50}$$

Upper bound on $\mathcal{L}^{S_2}(\widetilde{Y})$: For any derived $\widetilde{Y}^* \in \mathcal{P}(\widehat{Y})$ that is 0-discriminatory, using (49) and the identical arguments as that of Lemma 15, we have the following:

$$\mathbb{P}(\Gamma^{S_2}(\widetilde{Y}^*) > \widetilde{\alpha}_n) \overset{(a)}{\leq} \mathbb{P}\left(\Gamma^{S_2}(\widetilde{Y}^*) > 2 \max_{ya} \sqrt{\frac{2\log 64/\delta}{n\mathbf{P}_{ya}}}\right) \overset{(b)}{\leq} \frac{\delta}{4},$$

where (a) follows from the condition on $\widetilde{\alpha}_n$ and (b) from Lemma 16.

From the optimality of $\mathcal{L}^{S_2}(\widetilde{Y})$ and Lemma 16, $\forall \widetilde{Y}^* \in \mathcal{Q}(\mathcal{L}(\widetilde{Y}^*), 0) \cap \mathcal{P}(\widehat{Y})$ we have,

$$\mathcal{L}^{S_2}(\widetilde{Y}) \leq_{\delta/4} \mathcal{L}^{S_2}(\widetilde{Y}^*) \leq_{\delta/4} \mathcal{L}(\widetilde{Y}^*) + C_1 \sqrt{\frac{\log 8/\delta}{n}}. \tag{51}$$

Upper bound on $\mathcal{L}(\widetilde{Y}^*)$: The rest of the proof involves obtaining an upper bound for $\mathcal{L}(\widetilde{Y}^*)$ using Lemma 9. Recall from Lemma 9 that if h is an α-discriminatory binary predictor, the optimum non-discriminatory derived predictor $\widetilde{Y}^*(h)$ given by (3) satisfies, $\widetilde{Y}^*(h) \in \mathcal{Q}(\mathcal{L}(h) + \alpha, 0)$.

Thus, for $h = \widehat{Y}$, a derived non-discriminatory predictor $\widetilde{Y}^*(\widehat{Y})$ obtained from (3) satisfies,

$$\mathcal{L}(\widetilde{Y}^*(\widehat{Y})) \overset{(a)}{\leq_{\delta}} \mathcal{L}(Y^*) + \alpha_n + C_3 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathbf{P}_{ya}}}, \tag{52}$$

where in (a) $Y^* \in \mathcal{H}$ is any non-discriminatory predictor from the original hypothesis class $\mathcal{H}$ and the inequality follows from combining Lemma 8 and Lemma 9 as with probability atleast $1 - \delta$, $\widehat{Y}$ is atmost $\alpha = \Gamma(\widehat{Y}) \leq \alpha_n + C_2 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathbf{P}_{ya}}}$ discriminatory.

Combining (50), (51), and (52), with probability atleast $1 - 2\delta$,

$$\mathcal{L}(\widetilde{Y}) \leq \mathcal{L}(Y^*) + \alpha_n + C_1 \max_{ya} \sqrt{\frac{VC(\mathcal{H}) + \log 1/\delta}{n\mathbf{P}_{ya}}}, \text{ and}$$

$$\Gamma(\widetilde{Y}) \leq \widetilde{\alpha}_n + C_2 \max_{ya} \sqrt{\frac{\log 1/\delta}{n\mathbf{P}_{ya}}} \tag{53}$$

Theorem 7 follows from appropriate choice of $\alpha_n, \widetilde{\alpha}_n$ and rescaling δ. ∎

C.4. Proof of Theorem 10

Let the marginal distribution over (A, Y) be given by $p = \min_{a,y} \mathbb{P}(A = a, Y = y)$. Since the definiton of fairness is invariant to re-labelling of A, Y, assume without loss of generality that p corresponds to $A = 1, Y = 1$. For $\alpha \in (0, 1/2)$, the distribution $\mathcal{D}$ over $(X, A, Y) \in \{0, 1\}^n \times \{0, 1\} \times \{0, 1\}$ is described by

$$
\begin{aligned}
\mathbb{P}(X_1 = y \mid Y = y) &= 1 - \alpha \\
\mathbb{P}(X_i = 0 \mid Y = 0, A = 0) &= 1 && \text{for } i = 2, 3, ..., n \\
\mathbb{P}(X_i = 1 \mid Y = 1, A = 0) &= 1 && \text{for } i = 2, 3, ..., n \\
\mathbb{P}(X_i = 0 \mid Y = 0, A = 1) &= 1 && \text{for } i = 2, 3, ..., n \\
\mathbb{P}(X_i = 1 \mid Y = 1, A = 1) &= 1 - \alpha && \text{for } i = 2, 3, ..., n
\end{aligned}
\tag{54}
$$

Consider the following hypothesis class $\mathcal{H} = \{h_i\}_{i=1}^n$ with $h_i(X, A) = X_i$. The hypothesis h_1 has 0-1 loss $\mathcal{L}_{01}(h_1) = \mathbb{P}(X_1 \neq Y) = \alpha$ and is exactly non-discriminatory since $X_1 \perp A \mid Y$ by construction. For every other $i = 2, 3, ..., n$, the 0-1 loss of h_i is the same:

$$
\mathcal{L}_{01}(h_i) = \sum_y \sum_a \mathbb{P}(X_i = 1 - y \mid Y = y, A = a)\mathbb{P}(Y = y, A = a) = p\alpha
\tag{55}
$$

however, for these hypotheses $|\mathbb{P}(h_i = 1 | Y = 1, A = 1) - \mathbb{P}(h_i = 1 | Y = 1, A = 0)| = \alpha$ so h_i is α-discriminatory.

We will now show that on a sample S of size m, the empirical risk minimizer subject to an approximate non-discrimination constraint, $\widehat{h}$, will be h_i for $i \neq 1$ with probability 0.5. Hence, the first step alone cannot assure with probability better than 0.5 a classifier that is better than α-discriminatory.

First, we note that the predictions of h_i and h_j are independent for $i \neq j$ since X_i and X_j are independent. Therefore, the number of errors made by each classifier h_i on the sample S are independent and

$$
\mathbb{P}(\mathcal{L}_{01}^S(h_1) = 0) = (1 - \alpha)^m \qquad \mathbb{P}(\mathcal{L}_{01}^S(h_i) = 0) = (1 - p\alpha)^m \quad \text{for } i = 2, 3, ..., n
\tag{56}
$$

Since a classifier that makes zero errors on S is automatically non-discriminatory on S, if h_1 makes at least 1 mistake on S and some h_i does not make any errors, then h_1 will be the optimum of (10). This event occurs with probability:

$$
\mathbb{P}\left(\mathcal{L}_{01}^S(h_1) > 0 \wedge \exists i \, \mathcal{L}_{01}^S(h_i) = 0\right) = (1 - (1 - \alpha)^m)\left(1 - \mathbb{P}\left(\forall i > 1 \, \mathcal{L}_{01}^S(h_i) > 0\right)\right)
\tag{57}
$$

$$
= (1 - (1 - \alpha)^m)\left(1 - \prod_{i=2}^n (1 - (1 - p\alpha)^m)\right)
\tag{58}
$$

$$
= (1 - (1 - \alpha)^m)\left(1 - (1 - (1 - p\alpha)^m)^{n-1}\right)
\tag{59}
$$

From here, we use that

$$
\forall k \in \mathbb{N} \, \forall x \in [0, 1] \quad (1 - x)^k \leq \frac{1}{1 + kx}
\tag{60}
$$

Thus, since $1 - p\alpha, \alpha \in [0, 1]$:

$$\mathbb{P}\left(\mathcal{L}_{01}^S(h_1) > 0 \wedge \exists i \ \mathcal{L}_{01}^S(h_i) = 0\right) = (1 - (1 - \alpha)^m)\left(1 - (1 - (1 - p\alpha)^m)^{n-1}\right) \tag{61}$$

$$\geq \left(1 - \frac{1}{1 + m\alpha}\right)\left(1 - \frac{1}{1 + (n-1)(1 - p\alpha)^m}\right) \tag{62}$$

$$= \frac{m\alpha}{1 + m\alpha} - \left(1 - \frac{1}{1 + m\alpha}\right)\frac{1}{1 + (n-1)(1 - p\alpha)^m} \tag{63}$$

This expression is is greater than $1/2$ if the first term is at least $2/3$ and the second term is at most $1/6$. Thus

$$\frac{m\alpha}{1 + m\alpha} \geq \frac{2}{3} \iff \alpha \geq \frac{2}{m} \tag{64}$$

and

$$\left(1 - \frac{1}{1 + m\alpha}\right)\frac{1}{1 + (n-1)(1 - p\alpha)^m} \leq \frac{1}{6}$$

$$\Longleftarrow \frac{1}{1 + (n-1)(1 - p\alpha)^m} \leq \frac{1}{6} \iff \log\frac{1}{1 - p\alpha} \leq \frac{\log\frac{n-1}{5}}{m} \tag{65}$$

Since $-\log(1 - x) < \frac{x}{1-x}$ for $x \in (0, 1)$, and $p \leq \frac{1}{4}$, the expression (63) is at least $1/2$ when

$$\frac{2}{m} \leq \alpha \leq \frac{3 \log\frac{n}{5}}{4pm} \tag{66}$$

Therefore, when $\alpha = \frac{3 \log\frac{n}{5}}{4pm}$, with probability 0.5 h_1 has non-zero error on S and a different predictor has zero error. We conclude that there exists a distribution and hypothesis class such that with probability 0.5, the hypothesis returned by the first step is $\frac{3 \log\frac{n}{5}}{4pm}$-discriminatory.

Appendix D. Proof of Theorem 11

Let $\mathcal{A}$ be an algorithm that takes as inputs a hypothesis class $\mathcal{H}$, a distribution $\widetilde{\mathcal{D}}$ over $(\widetilde{X}, \widetilde{A}, \widetilde{Y})$ with $\widetilde{A} \in \{0, 1\}$ and $\widetilde{Y} \in \{-1, +1\}$, an accuracy parameter $\epsilon > 0$, and a non-discrimination parameter $\alpha > 0$ and returns a predictor $f = \mathcal{A}(\widetilde{\mathcal{D}}, \epsilon, \alpha)$ such that with probability $1 - \zeta$

$$\mathcal{L}_{\widetilde{\mathcal{D}}}^{\text{hinge}}(f) \leq \min_{h \in \mathcal{H}_{0\text{-disc}}(\widetilde{\mathcal{D}})} \mathcal{L}_{\widetilde{\mathcal{D}}}^{\text{hinge}}(h) + \epsilon \tag{67}$$

$$\left|\mathbb{P}_{\widetilde{\mathcal{D}}}\left(f \geq 0 \mid \widetilde{Y} = y, \widetilde{A} = 0\right) - \mathbb{P}_{\widetilde{\mathcal{D}}}\left(f \geq 0 \mid \widetilde{Y} = y, \widetilde{A} = 1\right)\right| \leq \alpha \qquad \text{for } y = -1, +1$$

The possibly randomized predictor f need not be in the hypothesis class $\mathcal{H}$, but it is being compared against the best predictor in $\mathcal{H}$ whose sign is non-discriminatory.

We will show that such an algorithm can be used to improperly weakly learn HALFSPACE which, subject to the complexity assumption that refuting random K-XOR formulas is hard, was shown to be computationally hard by Daniely (2015). We conclude that $\mathcal{A}$ must be computationally hard to compute.

The HALFSPACE problem is to take a distribution $\mathcal{D}$ over (X, Y) with $X \in \mathbb{R}^d$ and $Y \in \{-1, +1\}$, and find the linear predictor

$$h^*(x) = \text{sign}(w^{*T}x) \qquad \text{where} \qquad w^* = \underset{w \in \mathbb{R}^d}{\text{argmin}} \ \underset{x,y \sim \mathcal{D}}{\mathbb{E}}\left[\text{sign}(w^T x) \neq y\right] \tag{68}$$

The proof of hardenss of the HALFSPACE problem was shown using a distribution over the unit hypercube in d-dimensions, thus we will assume that $\mathcal{D}$ is a bounded distribution. We assume access to the distribution $\mathcal{D}$, knowledge of $\mathcal{L}_{\mathcal{D}}^{01}(h^*)$, and for now, access to the joint distribution of $(h^*(X), Y)$:

$$
\begin{aligned}
\eta_{--} &= \mathbb{P}_{\mathcal{D}}(h^*(X) = -1, Y = -1) & \eta_{-+} &= \mathbb{P}_{\mathcal{D}}(h^*(X) = -1, Y = +1) \\
\eta_{+-} &= \mathbb{P}_{\mathcal{D}}(h^*(X) = +1, Y = -1) & \eta_{++} &= \mathbb{P}_{\mathcal{D}}(h^*(X) = +1, Y = +1)
\end{aligned}
\tag{69}
$$

however, we will show later that it is not necessary to know the η's. Since it is always possible to get 0-1 loss at most $\frac{1}{2}$ with a HALFSPACE predictor, we assume that $\eta_{++} + \eta_{--} \geq \eta_{+-} + \eta_{-+} = \mathcal{L}_{\mathcal{D}}^{01}(h^*)$. From the distribution $\mathcal{D}$ we construct a new distribution $\widetilde{\mathcal{D}}$ over $(\widetilde{X}, \widetilde{A}, \widetilde{Y})$ with $\widetilde{X} \in \mathbb{R}^{d+1}$, $\widetilde{A} \in \{0, 1\}$, and $\widetilde{Y} \in \{-1, +1\}$ in the following manner:

$$
\begin{aligned}
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 0) &= 1 - \delta \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 1) &= \delta \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = -e_1, \widetilde{Y} = -1 \mid \widetilde{A} = 0) &= \eta_{--} \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = -e_1, \widetilde{Y} = +1 \mid \widetilde{A} = 0) &= \eta_{-+} \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = e_1, \widetilde{Y} = -1 \mid \widetilde{A} = 0) &= \eta_{+-} \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = e_1, \widetilde{Y} = +1 \mid \widetilde{A} = 0) &= \eta_{++} \\
\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = [0, x], \widetilde{Y} = y \mid \widetilde{A} = 1) &= \mathbb{P}_{\mathcal{D}}(X = x, Y = y) \qquad \forall x, y
\end{aligned}
\tag{70}
$$

where e_1 is the first standard basis vector in $\mathbb{R}^{d+1}$. In other words, when $\widetilde{A} = 1$ the distribution $\widetilde{\mathcal{D}}$ is identical to $\mathcal{D}$ besides a zero appended to the beginning of X. When $\widetilde{A} = 0$, $\widetilde{\mathcal{D}}$ is supported on two points $-e_1$ and e_1.

We will apply the algorithm $\mathcal{A}$ to the distribution $\widetilde{\mathcal{D}}$ with parameters ϵ and α to be determined later. In this case $\mathcal{H}$ is the class of linear predictors so the hinge loss of f on $\widetilde{\mathcal{D}}$ must be competitve with the hinge loss of the best linear predictor whose sign is non-discriminatory.

Using the following lemmas, we show that the output of $\mathcal{A}$ must have small hinge loss on $\widetilde{\mathcal{D}}$, that this output can then be modified so that it has small 0-1 loss on $\widetilde{\mathcal{D}}$, and finally that it can be further modified to achieve small 0-1 loss on $\mathcal{D}$. The proofs are deferred to the end of this discussion.

Lemma 17 *There exists a linear predictor h whose sign is 0-discrminatory such that*

$$
\mathcal{L}_{\widetilde{\mathcal{D}}}^{hinge}(h) = 2(1 - \delta)\mathcal{L}_{\mathcal{D}}^{01}(h^*) + 2\delta
$$

By Lemma 17 and the defintion of f from (67),

$$
\mathcal{L}_{\widetilde{\mathcal{D}}}^{hinge}(f) \leq \min_{h \in \mathcal{H}_{0\text{-disc}}(\widetilde{\mathcal{D}})} \mathcal{L}_{\widetilde{\mathcal{D}}}^{hinge}(h) + \epsilon \leq 2(1 - \delta)\mathcal{L}_{\mathcal{D}}^{01}(h^*) + 2\delta + \epsilon
\tag{71}
$$

and the sign of f is α-discriminatory. Next,

Lemma 18 *The predictor f can be efficiently modified to yield a new predictor f' whose sign is is α-discriminatory such that*

$$
\mathcal{L}_{\widetilde{\mathcal{D}}}^{01}(f') \leq (1 - \delta)\mathcal{L}_{\mathcal{D}}^{01}(h^*) + 2\delta + \epsilon
$$

Finally,

Lemma 19 *The predictor $f''(x) = f'([0,x], 1)$ achieves $\mathcal{L}_{\mathcal{D}}^{01}(f'') \leq \mathcal{L}_{\widetilde{\mathcal{D}}}^{01}(f') + \alpha(1 - \delta)$.*

The predictor f'' described in Lemma 19 thus has 0-1 loss on $\mathcal{D}$

$$\mathcal{L}_{\mathcal{D}}^{01}(f'') \leq (1 - \delta)(\mathcal{L}_{\mathcal{D}}^{01}(h^*) + \alpha) + 2\delta + \epsilon \tag{72}$$

Theorem 1.3 from Daniely (2015) proves that there is no algorithm running in time polynomial in the dimension d that can return a predictor achieving 0-1 error $\leq \frac{1}{2} - d^{-c}$ with high probability for a constant $c > 0$ for an arbitrary distribution, even with the knowledge that $\mathcal{L}_{\mathcal{D}}^{01}(h^*) \leq L^*$ for $L^* < 1/2$. Thus, $\mathcal{A}(\widetilde{D}, \epsilon, \alpha)$ cannot run in time polynomial in the dimension d for any parameters $\mathcal{L}_{\mathcal{D}}^{01}(h^*)$, ϵ, α, and δ such that (72) is greater than $\frac{1}{2} - d^{-c}$ for any $c > 0$. For $\epsilon, \alpha < \frac{1}{8}$, and setting $\delta = \frac{1}{16}$, (72) shows that

$$\mathcal{L}_{\mathcal{D}}^{01}(f'') \leq \frac{15}{16}\mathcal{L}_{\mathcal{D}}^{01}(h^*) + \frac{47}{128} \tag{73}$$

For any $L^* < \frac{1}{10}$ this is at most $\frac{1}{2} - \frac{1}{40}$ and does not depend on the dimension.

In this proof we assumed knowledge of the parameters η which describe the conditional error rates of h^*. If a polynomial time algorithm for $\mathcal{A}$ existed, then it would be possible to perform two-dimensional grid search over the η. Calls made to $\mathcal{A}$ with the incorrect values of η might result in very inaccurate or discriminatory predictors, but using an estimate of η up to $\mathcal{O}(\alpha)$ accuracy is sufficient to approximate the HALFSPACES solution using $\mathcal{A}$. Thus at most $\mathcal{O}(\log^2(1/\alpha))$ calls to the polynomial time algorithm would be needed. Therefore, in order for $\mathcal{A}$ to guarantee for an arbitrary $\widetilde{\mathcal{D}}$ that its output would have excess hinge loss at most $\frac{1}{8}$ and its sign would be at most $\frac{1}{8}$-discriminatory, it must run in time super-polynomial in the dimension in the worst case.

D.1. Deferred proofs

Proof [Proof of Lemma 17] Define $h(X, A) = \text{sign}\left(\begin{bmatrix} 1 \\ w^* \end{bmatrix}^T X\right)$. Then $h(-e_1, 0) = -1, h(e_1, 0) = 1$, and $h([0, x], 1) = h^*(x)$, where h^* is as defined in (68). Since the sign function is invariant to scaling, $\mathcal{L}_{\mathcal{D}}^{01}(h^*) = \mathcal{L}_{\mathcal{D}}^{01}(ch^*)$ for any $c > 0$.

Theorem 1.3 in Daniely (2015) involves a distribution $\mathcal{D}$ that is supported on the unit hypercube in $\mathbb{R}^d$, thus the predictor $\frac{h^*}{\|w^*\|_2 \sqrt{d}} \in [-1, 1]$ with probability 1, and has the same 0-1 loss as h^*.

By the definition of $\widehat{\mathcal{D}}$, η_{+-}, and η_{-+}:

$$\begin{aligned}
\mathbb{P}_{\widehat{\mathcal{D}}}(h \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 0) &= \frac{\eta_{+-}}{\mathbb{P}_{\widehat{\mathcal{D}}}(\widetilde{Y} = -1 \mid \widetilde{A} = 0)} \\
&= \frac{\eta_{+-}}{\mathbb{P}_{\mathcal{D}}(Y = -1)} \\
&= \mathbb{P}_{\mathcal{D}}(h^* \geq 0 \mid Y = -1) \\
&= \mathbb{P}_{\widehat{\mathcal{D}}}(h \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1)
\end{aligned} \tag{74}$$

$$\mathbb{P}_{\widetilde{\mathcal{D}}}(h < 0 \mid \widetilde{Y} = 1, \widetilde{A} = 0) = \frac{\eta_{-+}}{\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1 \mid \widetilde{A} = 0)}$$

$$= \frac{\eta_{-+}}{\mathbb{P}_{\mathcal{D}}(Y = 1)} \tag{75}$$

$$= \mathbb{P}_{\mathcal{D}}(h^* < 0 \mid Y = 1)$$

$$= \mathbb{P}_{\widetilde{\mathcal{D}}}(h < 0 \mid \widetilde{Y} = 1, \widetilde{A} = 1)$$

Therefore, h is 0-discriminatory at threshold 0. Also

$$
\begin{aligned}
\mathcal{L}_{\widetilde{\mathcal{D}}}^{\text{hinge}}(h) &= [1 + h(x_1)]_+ \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x_1, \widetilde{A} = 0, \widetilde{Y} = -1) \\
&\quad + [1 - h(x_0)]_+ \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x_0, \widetilde{A} = 0, \widetilde{Y} = 1) \\
&\quad + \int_X [1 + h(x)]_+ \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x, \widetilde{A} = 1, \widetilde{Y} = -1) dx \\
&\quad + \int_X [1 - h(x)]_+ \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x, \widetilde{A} = 1, \widetilde{Y} = 1) dx
\end{aligned}
\tag{76}
$$

$$
\begin{aligned}
&\le 2\eta_{+-} \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 0) + 2\eta_{-+} \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 0) \\
&\quad + 2\int_X \left(\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x, \widetilde{A} = 1, \widetilde{Y} = -1) + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{X} = x, \widetilde{A} = 1, \widetilde{Y} = 1) \right) dx
\end{aligned}
\tag{77}
$$

$$= 2\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 0)(\eta_{+-} + \eta_{-+}) + 2\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 1) \tag{78}$$

$$= 2(1 - \delta)(\eta_{-+} + \eta_{+-}) + 2\delta \tag{79}$$

∎

Proof [Proof of Lemma 18] Since only the sign of f is required to be α-discriminatory, we can modify the magnitude of its predictions without affecting its level of non-discrimination. Therefore, we first truncate the output of f to lie in the range $[-1, 1]$, which can only reduce the hinge loss.

Ignoring for the moment that the sign of f must be α-discriminatory, we would like to define f' so that $f'(-e_1, 0) = -1$ and $f'(e_1, 0) = 1$ with probability 1. In this case, the hinge loss when $\widetilde{A} = 0$ is exactly $2(\eta_{+-} + \eta_{-+})$. With that being said, any modification to f that changes the distribution of the *sign* of the predictor risks rendering it more than α-discriminatory. With this in mind, we construct f' such that

$$
\begin{aligned}
f'(-e_1, 0) &= \begin{cases} -1 & \text{w.p. } \mathbb{P}(f(-e_1, 0) < 0) \\ 0 & \text{w.p. } \mathbb{P}(f(-e_1, 0) \ge 0) \end{cases} \\
f'(e_1, 0) &= \begin{cases} 1 & \text{w.p. } \mathbb{P}(f(e_1, 0) \ge 0) \\ -0 & \text{w.p. } \mathbb{P}(f(e_1, 0) < 0) \end{cases} \\
f'([0, x], 1) &= f([0, x], 1)
\end{aligned}
\tag{80}
$$

where -0 is a negative number of arbitrarily small magnitude. Constructed this way, the distribution of the sign of f' conditioned on $\widetilde{A}$ is identical to that of f, meaning that the sign of f' is α-discriminatory.

The hinge loss of f' is an upper bound on the 0-1 loss, and in order to show that f' achieves small 0-1 loss, we will show that the hinge loss is a *loose* upper bound. The construction of

$\widetilde{\mathcal{D}}$ and the predictions of f' conditioned on $\widetilde{A} = 0$ creates a substantial gap between the losses.
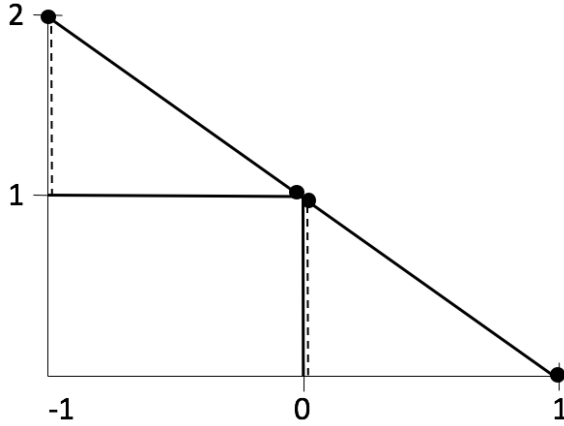
Notice that when f' makes a prediction of magnitude 1 that has the correct sign, both the hinge loss and the 0-1 loss evaluate to 0. Similarly, when f' makes a prediction of magnitude 0 with the incorrect sign, both losses are 1. Thus in each of these cases, the hinge loss is equivalent to the 0-1 loss.

However, if f' makes a prediction of magnitude 1 with the incorrect sign, the hinge loss is 2 but the 0-1 loss is only 1, and when f' makes a prediction of magnitude 0 with the correct sign, the hinge loss is 1 but the 0-1 loss is 0. Consequently, in each of these cases there is a gap of 1 between the hinge and 0-1 losses. Thus,

$$\mathbb{E}\left[\ell^{\text{hinge}}(f') - \ell^{01}(f') \,\Big|\, \widetilde{A} = 0\right] = \mathbb{P}\left(|f'| = 1, \text{sign}(f') \neq \widetilde{Y} \,\Big|\, \widetilde{A} = 0\right) \tag{81}$$
$$+ \mathbb{P}\left(|f'| = 0, \text{sign}(f') = \widetilde{Y} \,\Big|\, \widetilde{A} = 0\right)$$

Considering each term separately:

$$\mathbb{P}\left(|f'| = 1, \text{sign}(f') \neq \widetilde{Y} \,\Big|\, \widetilde{A} = 0\right)$$
$$= \mathbb{P}\left(f'(-e_1, 0) = -1, \widetilde{Y} = 1 \,\Big|\, \widetilde{X} = -e_1, \widetilde{A} = 0\right) \mathbb{P}\left(\widetilde{X} = -e_1 \,\Big|\, \widetilde{A} = 0\right) \tag{82}$$
$$+ \mathbb{P}\left(f'(e_1, 0) = 1, \widetilde{Y} = -1 \,\Big|\, \widetilde{X} = e_1, \widetilde{A} = 0\right) \mathbb{P}\left(\widetilde{X} = e_1 \,\Big|\, \widetilde{A} = 0\right)$$
$$= \mathbb{P}\left(f'(-e_1, 0) = -1\right) \mathbb{P}\left(\widetilde{X} = -e_1, \widetilde{Y} = 1 \,\Big|\, \widetilde{A} = 0\right) \tag{83}$$
$$+ \mathbb{P}\left(f'(e_1, 0) = 1\right) \mathbb{P}\left(\widetilde{X} = e_1, \widetilde{Y} = -1 \,\Big|\, \widetilde{A} = 0\right)$$
$$= \mathbb{P}\left(f(-e_1, 0) < 0\right) \eta_{-+} + \mathbb{P}\left(f(e_1, 0) \geq 0\right) \eta_{+-} \tag{84}$$

With (83) following from the fact that conditioned on $\widetilde{X}$ and $\widetilde{A} = 0$, $f'(\widetilde{X}, 0)$ is a random variable that is independent of the value of $\widetilde{Y}$. Similarly,

$$\mathbb{P}\left(|f'(\widetilde{X}, 0)| = 0, \operatorname{sign}(f'(\widetilde{X}, 0)) = \widetilde{Y} \mid \widetilde{A} = 0\right)$$

$$= \mathbb{P}\left(f'(-e_1, 0) = 0, \widetilde{Y} = 1 \mid \widetilde{X} = -e_1, \widetilde{A} = 0\right) \mathbb{P}\left(\widetilde{X} = -e_1 \mid \widetilde{A} = 0\right) \quad (85)$$

$$+ \mathbb{P}\left(f'(e_1, 0) = -0, \widetilde{Y} = -1 \mid \widetilde{X} = e_1, \widetilde{A} = 0\right) \mathbb{P}\left(\widetilde{X} = e_1 \mid \widetilde{A} = 0\right)$$

$$= \mathbb{P}\left(f'(-e_1, 0) = 0\right) \mathbb{P}\left(\widetilde{X} = -e_1, \widetilde{Y} = 1 \mid \widetilde{A} = 0\right) \quad (86)$$

$$+ \mathbb{P}\left(f'(e_1, 0) = -0\right) \mathbb{P}\left(\widetilde{X} = e_1, \widetilde{Y} = -1 \mid \widetilde{A} = 0\right)$$

$$= \mathbb{P}\left(f(-e_1, 0) \geq 0\right) \eta_{-+} + \mathbb{P}\left(f(e_1, 0) < 0\right) \eta_{+-} \quad (87)$$

Combining (81) with (84) and (87), we see that

$$\mathbb{E}\left[\ell^{\mathrm{hinge}}(f') - \ell^{01}(f') \mid \widetilde{A} = 0\right] = \mathbb{P}\left(f(-e_1, 0) < 0\right) \eta_{-+} + \mathbb{P}\left(f(e_1, 0) \geq 0\right) \eta_{+-} \quad (88)$$

$$+ \mathbb{P}\left(f(-e_1, 0) \geq 0\right) \eta_{-+} + \mathbb{P}\left(f(e_1, 0) < 0\right) \eta_{+-}$$

$$= \eta_{-+} + \eta_{+-} \quad (89)$$

The 0-1 loss of f' can be decomposed as

$$\mathcal{L}_{\widetilde{\mathcal{D}}}^{01}(f') = \mathbb{P}(\widetilde{A} = 0)\mathbb{E}\left[\ell^{01}(f) \mid \widetilde{A} = 0\right] + \mathbb{P}(\widetilde{A} = 1)\mathbb{E}\left[\ell^{01}(f) \mid \widetilde{A} = 1\right] \quad (90)$$

By (89),

$$\mathbb{P}(\widetilde{A} = 0)\mathbb{E}\left[\ell^{01}(f) \mid \widetilde{A} = 0\right] = (1 - \delta)\left(\mathbb{E}\left[\ell^{\mathrm{hinge}}(f) \mid \widetilde{A} = 0\right] - \eta_{-+} - \eta_{+-}\right) \quad (91)$$

and since the hinge loss is always an upper bound on the 0-1 loss

$$\mathbb{P}(\widetilde{A} = 1)\mathbb{E}\left[\ell^{01}(f) \mid \widetilde{A} = 1\right] \leq \delta\mathbb{E}\left[\ell^{\mathrm{hinge}}(f) \mid \widetilde{A} = 1\right] \quad (92)$$

From Lemma 17, the hinge loss of f' (which is at most the hinge loss of f) is upper bounded by $2(1 - \delta)(\eta_{-+} + \eta_{+-}) + 2\delta + \epsilon$. Thus,

$$\mathcal{L}_{\widetilde{\mathcal{D}}}^{01}(f') \leq (1 - \delta)\left(\mathbb{E}\left[\ell^{\mathrm{hinge}}(f) \mid \widetilde{A} = 0\right] - \eta_{-+} - \eta_{+-}\right) + \delta\mathbb{E}\left[\ell^{\mathrm{hinge}}(f) \mid \widetilde{A} = 1\right] \quad (93)$$

$$= \mathcal{L}_{\widetilde{\mathcal{D}}}^{\mathrm{hinge}}(f') - (1 - \delta)(\eta_{-+} + \eta_{+-}) \quad (94)$$

$$\leq (1 - \delta)(\eta_{-+} + \eta_{+-}) + 2\delta + \epsilon \quad (95)$$

$\blacksquare$

Proof [Proof of Lemma 19] Because f' is α-discriminatory at threshold 0

$$\left|\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 0) - \mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1)\right| \leq \alpha$$

$$\left|\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 0) - \mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1)\right| \leq \alpha \quad (96)$$

Let $f''(x) = f'([0, x], 1)$, then

$$
\begin{aligned}
\mathcal{L}_{\widetilde{\mathcal{D}}}^{01}(h) &= \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 0)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 0) &&(97)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 0)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 0)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 1)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 1)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1)\\
&\geq \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 0)\left(\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1) - \alpha\right) &&(98)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 0)\left(\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1) - \alpha\right)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 1)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 1)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1)\\
&= \left(\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 0) + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 1)\right)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1) &&(99)\\
&\quad + \left(\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 0) + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 1)\right)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1)\\
&\quad + \left(\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1, \widetilde{A} = 0) + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1, \quad \widetilde{A} = 0)\right)(-\alpha)\\
&= \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = -1)\mathbb{P}_{\widetilde{\mathcal{D}}}(f' \geq 0 \mid \widetilde{Y} = -1, \widetilde{A} = 1) &&(100)\\
&\quad + \mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{Y} = 1) \ \ \mathbb{P}_{\widetilde{\mathcal{D}}}(f' < 0 \mid \widetilde{Y} = 1, \quad \widetilde{A} = 1) - \alpha\mathbb{P}_{\widetilde{\mathcal{D}}}(\widetilde{A} = 0) &&(101)\\
&= \mathbb{P}_{\mathcal{D}}(Y = -1)\mathbb{P}_{\mathcal{D}}(f'' \geq 0 \mid Y = -1) &&(102)\\
&\quad + \mathbb{P}_{\mathcal{D}}(Y = 1)\mathbb{P}_{\mathcal{D}}(f'' < 0 \mid Y = 1) - \alpha(1 - \delta)\\
&= \mathcal{L}_{\mathcal{D}}^{01}(f'') - \alpha(1 - \delta) &&(103)
\end{aligned}
$$

The lemma follows immediately. ∎

Appendix E. Proofs for Section 6 - Relaxing non-discrimination

We use $\Sigma_{\cdot}$ to denote covariances involving a vector and we reserve $\sigma_{\cdot}$ for scalar covariances. We start with recalling some facts about Gaussian random variables.

Proposition 20 *If (U, V, W) are jointly Gaussian, then:*

- ***Conditional expectation*** $\mathbb{E}[U|V]$ *is linear in V and is given by:*

$$\mathbb{E}[U|V] = \mathbb{E}[U] + \Sigma_{U,V}\Sigma_V^{-1}(V - \mathbb{E}[V]).$$

- ***Conditional covariance*** $\Sigma_{U,V|W}$ *does not depend on W and it is always equal to:*

$$\Sigma_{U,V|W} = \Sigma_{U,V} - \Sigma_{U,W}\Sigma_W^{-1}\Sigma_{W,V}.$$

E.1. Proof of Theorem 13

First, let us show the second claim. Let R be any linear predictor. By linearity, it follows that (R, A, Y) are jointly Gaussian. By the conditional covariance formula, we have:

$$\sigma_{RA|Y} = \sigma_{RA} - \sigma_{RY}\sigma_{YA}/\sigma_Y^2.$$

If R satisfies equalized correlations, then the right-hand side here is 0. It follows that R and A are uncorrelated conditionally on Y. But since they are jointly Gaussian, they are also independent conditionally on Y. Therefore R also satisfies equalized odds non-discrimination. The converse also holds: if R satisfies equalized odds, then R and A are uncorrelated given Y, and therefore equalized correlations is satified.

Now let us move back to the main claim. Assume, without loss of generality, that all variables are centered. Let us first find the optimal (a priori not necessarily linear) predictor that satisfies the relaxed second-moment non-discrimination criterion. In particular, the Lagrangian to minimize may be written as:

$$\mathbb{E}[(R - Y)^2] - \lambda\left(\sigma_Y^2\mathbb{E}[RA] - \sigma_{AY}\mathbb{E}[RY]\right),$$

But just like in the unconstrained least squares problem, we may apply the law of total expectatons to condition the loss and the R- terms in the second-moment non-discrimination constraint to be conditioned on X and A. Thus the optimum is achieved for each X, A by minimizing the following Lagrangian:

$$\mathbb{E}[(R - Y)^2|X, A] - \lambda\left(\sigma_Y^2 AR - \sigma_{AY}\mathbb{E}[Y|X, A]R\right),$$

or equivalently

$$R^2 - 2\mathbb{E}[Y|X, A]R - \lambda\left(\sigma_Y^2 AR - \sigma_{AY}\mathbb{E}[Y|X, A]R\right).$$

It follows that the optimal R is a linear function of A, $\mathbb{E}[Y|X, A]$ and λ. λ is determined over the statistics of the problem, and therefore it is a constant that does not depend on specific values of X and A, and $\mathbb{E}[Y|X, A]$ in the Gaussian setting is linear. It thus follows that the optimum, let's call it $R_\circ$, is a linear function of X and A. It minimizes the expected square loss subject to a relaxed non-discrimination criterion, therefore it is not larger than the optimizer under the stricter constraint. Conversely, by linearity it does also satisfy the stricter constraint, and is thus no smaller than the optimizer under that constraint (recall that we didn't start out by imposing linearity). Therefore $R_\circ$ is precisely the squared loss optimal equalized odds non-discriminatory predictor.

E.2. Optimal equalized correlations linear predictor

We write the proofs more generally for a vector-valued protected attribute A, and the scalar case follows directly. In this case v is a matrix and α is a vector mixing the columns of v (so the correction has the term $v\alpha$). First note that condition (13) translates into a linear constraint on w in the least-squares problem of (15). Using the bilinearity of the covariance, this constraint is equivalent to:

$$w^T\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],A} - w^T\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],Y}\,\Sigma_{Y,A}/\sigma_Y^2 \equiv w^T v = 0.$$

We can now write the cost function with a vector of Lagrange multipliers:

$$J(w, \lambda) = \mathbb{E}[(Y - w^T\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right])^2] + w^T v\lambda,$$

whose gradient is

$$\nabla_w J = 2\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} w - 2\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],Y} + v\lambda.$$

Setting this to zero, the optimality conditions give us the claimed functional form. The vector α can then be obtained by enforcing the constraint:

$$w^T v = \left(\Sigma_{Y,\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} - \alpha^T v^T\right)\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} v = 0$$

and thus

$$v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} v\alpha = v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],Y}.$$

E.3. Proof of Theorem 14

First, let us rewrite $R^\star$ as:

$$\begin{aligned} R^\star &= w^{\star T}\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right] \\ &= \widehat{w}^T\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right] - \alpha^T v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right] \\ &= \widehat{R} - \alpha^T v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right] \end{aligned}$$

Next, recall that

$$v = \Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],A} - \Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],Y}\Sigma_{Y,A}/\sigma_Y^2,$$

and since

$$\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} = \begin{bmatrix}\Sigma_{X,\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\\\Sigma_{A,\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\end{bmatrix} \Rightarrow \Sigma_{A,\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} = \left[0_{\dim(A)\times\dim(X)}\ I_{\dim(A)}\right],$$

we have

$$v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} = \left[0\ I\right] - \Sigma_{A,Y}\widehat{w}^T/\sigma_Y^2.$$

Therefore

$$v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right] = A - \frac{\widehat{R}}{\sigma_Y^2}\Sigma_{A,Y}$$

and can be derived from $(\widehat{R}, A, Y)$. Then multiplying from the right by v and using the bilinearity of the covariance, we get the terms in α:

$$v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]} v = \Sigma_{A,A} - \tfrac{1}{\sigma_Y^2}\Sigma_{A,Y}\Sigma_{Y,A} - \tfrac{1}{\sigma_Y^2}\Sigma_{A,Y}\left(\Sigma_{\widehat{R},A} - \tfrac{1}{\sigma_Y^2}\Sigma_{\widehat{R},Y}\Sigma_{Y,A}\right),$$

$$\text{and } v^T\Sigma^{-1}_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right]}\Sigma_{\left[\begin{smallmatrix}X\\A\end{smallmatrix}\right],Y} = \Sigma_{A,Y} - \tfrac{1}{\sigma_Y^2}\Sigma_{A,Y}\Sigma_{\widehat{R},Y}.$$

This shows that α also derives from $(\widehat{R}, A, Y)$ as stated. To simplify the expression of α to the one claimed, note that we have $\Sigma_{\widehat{R},A} = \Sigma_{Y,A}$. This is because $\widehat{R}$ is the squared loss optimal linear predictor of Y given A and X, and thus $\widehat{R} - Y$ is uncorrelated with any linear function of A and X, and in particular A. This completes the proof.