The Impact of Split Classifiers on Group Fairness

Hao Wang*, Hsiang Hsu*, Mario Diaz†, and Flavio P. Calmon*
*Harvard University, {hao_wang,hsianghsu}@g.harvard.edu; flavio@seas.harvard.edu
†Universidad Nacional Autónoma de México, mario.diaz@sigma.iimas.unam.mx

2021 IEEE International Symposium on Information Theory (ISIT) | 978-1-5386-8209-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/ISIT45174.2021.9517723

***Abstract*—Disparate treatment occurs when a machine learning model produces different decisions for groups of individuals based on a sensitive attribute (e.g., age, sex). In domains where prediction accuracy is paramount, it could potentially be acceptable to fit a model which exhibits disparate treatment. To evaluate the effect of disparate treatment, we compare the performance of split classifiers (i.e., classifiers trained and deployed separately on each group) with group-blind classifiers (i.e., classifiers which do not use a sensitive attribute). We introduce the benefit-of-splitting for quantifying the performance improvement by splitting classifiers when the underlying data distribution is known. Computing the benefit-of-splitting directly from its definition involves solving optimization problems over an infinite-dimensional functional space. Under different performance measures, we (i) prove an equivalent expression for the benefit-of-splitting which can be efficiently computed by solving small-scale convex programs; (ii) provide sharp upper and lower bounds for the benefit-of-splitting which reveal precise conditions where a group-blind classifier will always suffer from a non-trivial performance gap from the split classifiers.**

A full version of this paper is accessible at [1].

I. Introduction

A machine learning (ML) model exhibits *disparate treatment* [2] if it treats individuals differently based on a sensitive attribute (e.g., age, sex). In applications such as hiring, the existence of disparate treatment can be illegal [3]. However, in settings such as healthcare, it can be legal and ethical to fit a model which presents disparate treatment in order to improve prediction accuracy. For example, the Equal Credit Opportunity Act (ECOA) permits a creditor to use an applicant's age and income for analyzing credit, as long as such information is used in a fair manner (see 12 CFR §1002.6(b)(2) in [4]).

The role of a sensitive attribute in fair classification can be understood through several metrics and principles. When a ML model is deployed in practice, fairness can be quantified in terms of the performance disparity *conditioned* on a sensitive attribute, such as statistical parity [5] and equalized odds [6]. In domains where the goal is to predict accurately (e.g., medical diagnostics), *non-maleficence* (i.e., "do no harm") and *beneficence* (i.e., "do good") [7] become more appropriate moral principles for fairness [8–10]. Accordingly, a ML model should avoid the causation of harm and be as accurate as possible on each protected group.

The relationship between achieving the above-mentioned principles and allowing a classifier to exhibit disparate treatment is complex. On the one hand, using a *group-blind classifier* (i.e., a classifier that does not use the sensitive attribute as an input feature) may cause harm unintentionally since model performance relies on the distribution of the input data [8, 11–13]. This probability distribution can vary significantly conditioned on a sensitive attribute due to, for example, inherent differences between groups [11], differences in labeling [14], and differences in sampling [15]. On the other hand, training a separate classifier for each protected group—a setting we refer to as *splitting classifiers*—does not necessarily guarantee non-maleficence when sample size is limited [16]: groups with insufficient samples may incur a high generalization error and suffer from overfitting.

We consider two questions that are central to understanding non-maleficence and beneficence through the use of a sensitive attribute by a ML model:

1) When is it beneficial to split classifiers in terms of model performance?
2) When splitting is beneficial, how much do the split classifiers outperform a group-blind classifier?

First, we show that when the underlying distribution is known—or, equivalently, an arbitrarily large number of samples are available—splitting *never harms* any group in terms of average performance metrics. Thus, splitting will naturally follow the non-maleficence principle in the large-sample regime. Second, we introduce a notion called the *benefit-of-splitting* which measures the performance improvement by splitting classifiers compared to using a group-blind classifier across all groups. The benefit-of-splitting is also an information-theoretic quantity as it only relies on the underlying data distribution rather than number of samples or hypothesis class.

The definition of the benefit-of-splitting involves a model performance measure and, hence, we divide our analyses into two parts based on different choices of this measure. In Section III, we quantify model performance in terms of standard loss functions (e.g., ℓ_1 loss). For the benefit-of-splitting under these loss functions, we provide sharp upper and lower bounds (Theorem 1) that capture when splitting classifiers benefits model performance the most. These bounds indicate two factors (see Figure 1 for an illustration) which are central to the benefit-of-splitting: (i) disagreement between labeling functions[1], (ii) similarity between unlabeled distributions[1].

In Section IV, we consider false error rate as a performance measure since in applications such as medical diagnostics, high

[1] We borrow the terms "labeling function" and "unlabeled distribution" from the domain adaptation literature [17, 18]. The unlabeled distribution is a (marginal) probability distribution of the unlabeled data. The labeling function takes a data point as an input and produces the probability of its binary label being 1. Furthermore, the labeling function can be viewed as a "channel" (i.e., conditional distribution) in the information theory parlance. The formal definitions are given in Section I-A.

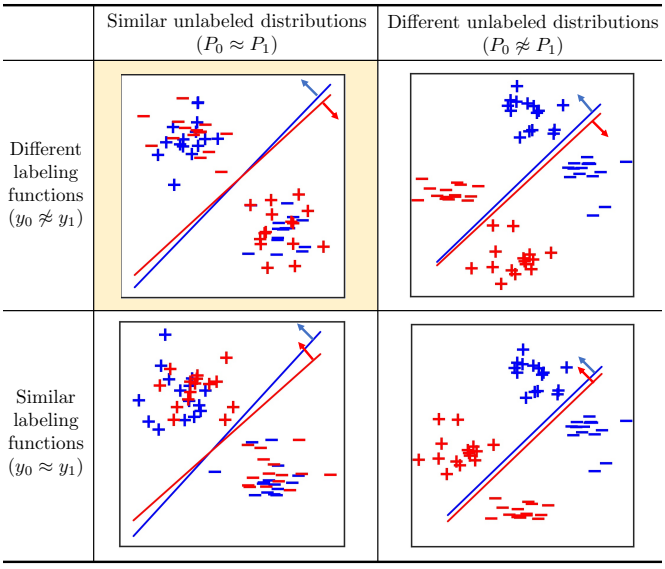

Fig. 1. The taxonomy of splitting based on two different factors. Samples from two groups are depicted in red and blue, respectively, and their labels are represented by $+$, $-$. Each group's labeling function is shown with the corresponding color and the arrows indicate the regions where the points are labeled as $+$. Information-theoretically, splitting classifiers benefits model performance the most if the labeling functions are different and the unlabeled distributions are similar (yellow region).

false error rate could result in unintentional harm [19]. Under this metric, computing the benefit-of-splitting directly from its definition may at first seem intractable since it involves an optimization over an infinite-dimensional functional space. Nonetheless, we prove that the benefit-of-splitting under false error rate has an equivalent, dual expression (Theorem 2) which only requires solving two small-scale convex programs. Furthermore, the objective functions of these convex programs have closed-form supergradients (Proposition 1). Combining these two results leads to an efficient procedure for computing the benefit-of-splitting.

The proof techniques of this paper are based on fundamental tools found in statistics, such as Brown-Low's two-points lower bound [20], and methods in convex analysis, such as Ky Fan's min-max theorem [21]. These tools are widely used in applications such as non-parametric estimation [22], and are useful for analyzing the min-max risk in statistical settings [23–27]. Furthermore, the factors that we provide for understanding the effect of splitting classifiers are inspired by the necessary and sufficient conditions of domain adaptation learnability in Ben-David *et al.* [28].

A. Notation and Definitions

Consider a binary classification task (e.g., detecting pneumonia from X-rays) where the goal is to learn a probabilistic classifier $h : \mathcal{X} \to [0,1]$ that predicts a label (e.g., presence of pneumonia) $Y \in \{0,1\}$ using input features (e.g., chest X-rays) $X \in \mathcal{X}$. We assume there is an additional binary sensitive attribute (e.g., sex) $S \in \{0,1\}$ that does not belong to the input features X. We denote the unlabeled probability distributions of input features conditioned on the sensitive attribute by

$$P_0 \triangleq P_{X|S=0}, \quad P_1 \triangleq P_{X|S=1}.$$

The labeling functions of the two groups are denoted by

$$y_0(x) \triangleq P_{Y|X,S}(1|x,0), \quad y_1(x) \triangleq P_{Y|X,S}(1|x,1).$$

In order to measure the difference between two unlabeled distributions, we recall Csiszár's f-divergence [29]. Let $f : (0,\infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and P_0, P_1 be two probability distributions over $\mathcal{X}$. The f-divergence between P_0 and P_1 is defined by

$$\mathrm{D}_f(P_0 \| P_1) \triangleq \int_{\mathcal{X}} f\left(\frac{\mathrm{d}P_0}{\mathrm{d}P_1}\right) \mathrm{d}P_1. \tag{1}$$

II. The Benefit-of-Splitting

We study the impact of disparate treatment by comparing the performance between optimal group-blind and split classifiers. First, we illustrate the difference between group-blind and split classifiers through the example of logistic regressions:

- a group-blind classifier does not use a sensitive attribute as an input: $h(x) = \mathsf{logistic}(w^T x)$ where $\mathsf{logistic}(t) \triangleq 1/(1+\exp(-t))$ for $t \in \mathbb{R}$;
- split classifiers are a set of classifiers trained and deployed separately on each group: $h_s(x) = \mathsf{logistic}(w_s^T x)$ for $s \in \{0,1\}$.

We measure the performance of both group-blind and split classifiers in terms of the *disadvantaged group* (i.e., the group with worst performance). For a given performance measure $L_s(\cdot)$ (higher values indicate a worse performance), the performance of a group-blind classifier h and a set of split classifiers $\{h_s\}_{s\in\{0,1\}}$, respectively, is measured by

$$\max_{s\in\{0,1\}} L_s(h) \quad \text{and} \quad \max_{s\in\{0,1\}} L_s(h_s).$$

Consequently, the optimal group-blind and split classifiers achieve the performance

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) \quad \text{and} \quad \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h).$$

Next, we introduce the benefit-of-splitting to quantify the effect of splitting classifiers compared to using a group-blind classifier.

Definition 1. For a given performance measure $L_s(\cdot)$ with $s \in \{0,1\}$, we define the benefit-of-splitting as

$$\epsilon_{\text{split}} \triangleq \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) - \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h), \tag{2}$$

where the infimum is taken over all (measurable) functions.

The benefit-of-splitting is the difference between the performance of the optimal group-blind and split classifiers. In other words, if h^* and $\{h_s^*\}_{s\in\{0,1\}}$ are optimal group-blind and split classifiers respectively, i.e.,

$$h^* \in \operatorname*{argmin}_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h),$$
$$h_s^* \in \operatorname*{argmin}_{h:\mathcal{X}\to[0,1]} L_s(h) \quad s \in \{0,1\},$$

the benefit-of-splitting can be equivalently expressed as

$$\epsilon_{\text{split}} = \max_{s\in\{0,1\}} L_s(h^*) - \max_{s\in\{0,1\}} L_s(h_s^*). \tag{3}$$

By the optimality of h_s^* and the max-min inequality, we have $L_s(h^*) \geq L_s(h_s^*)$ for $s \in \{0,1\}$ and $\epsilon_{\text{split}} \geq 0$ which implies that, information-theoretically, using a separate classifier on each group will never diminish model performance compared to using a group-blind classifier. A natural question is: how much performance improvement does splitting classifiers bring? Before answering this question, we specify performance measures of interest and present the benefit-of-splitting under these performance measures.

A. Loss Reduction by Splitting

The first type of performance measures contains standard loss functions which quantify the disagreement between the labeling function y_s and the probabilistic classifier h. We recast the benefit-of-splitting under these loss functions below.

Definition 2. The ℓ_1-benefit-of-splitting $\epsilon_{\text{split},1}$ is the benefit-of-splitting in Definition 1 with the performance measure:

$$L_s(h) = \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right].$$

The ℓ_2-benefit-of-splitting $\epsilon_{\text{split},2}$ is the benefit-of-splitting in Definition 1 with the performance measure:

$$L_s(h) = \mathbb{E}\left[(h(X) - y_s(X))^2 \mid S = s\right].$$

The KL-benefit-of-splitting $\epsilon_{\text{split},\mathsf{KL}}$ is the benefit-of-splitting in Definition 1 with the performance measure:

$$L_s(h) = \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}(y_s(X)\|h(X)) \mid S = s\right],$$

where $\mathsf{D}_{\mathsf{KL}}(p\|q) \triangleq p\log(p/q) + (1-p)\log((1-p)/(1-q))$ for $p, q \in [0,1]$.

B. False Error Rate Reduction by Splitting

Now we use the false error rate (FER) as a performance measure. The false error rate of a classifier is the maximum between (generalized) false positive rate and (generalized) false negative rate [30].

Definition 3. The FER-benefit-of-splitting $\epsilon_{\text{split},\mathsf{FER}}$ is the benefit-of-splitting in Definition 1 with the performance measure $L_s(h)$ being equal to

$$\max\left\{\mathbb{E}\left[h(X)|Y=0, S=s\right], \mathbb{E}\left[1-h(X)|Y=1, S=s\right]\right\}.$$

a) Connection with equalized odds: Equalized odds [6] is a commonly used group fairness measure that requires different groups to have the same false positive and false negative rates. However, imposing equalized odds constraints may lead to a performance reduction in classification [31–34]. In contrast, the benefit-of-splitting aims to capture the principles of non-maleficence and beneficence: classifiers should avoid the causation of harm and achieve the best performance on each group. By taking the optimal group-blind classifier as a baseline approach, this may allow split classifiers to potentially exhibit performance disparities between groups—as long as the split classifiers do not perform worse than the baseline approach and are as accurate as possible.

III. The Taxonomy of Splitting

In this section, we analyze the loss reduction by splitting classifiers compared to using a group-blind classifier. We achieve this goal by upper and lower bounding the benefit-of-splitting under different loss functions (Definition 2). These bounds reveal factors which could impact the effect of splitting classifiers and lead to a taxonomy of splitting, i.e., a characterization of when splitting benefits model performance the most or splitting does not bring much benefit.

Theorem 1. *The ℓ_1-benefit-of-splitting $\epsilon_{split,1}$ can be upper bounded by*

$$\min\left\{\min_{s\in\{0,1\}} \mathsf{D}_2(y_1, y_0|s)\cdot\sqrt{1-\mathsf{D}_{\mathsf{TV}}(P_0\|P_1)}, \frac{1}{2}\max_{s\in\{0,1\}}\mathsf{D}_1(y_1,y_0|s)\right\}$$

and lower bounded by

$$\frac{1}{2}\max_{s\in\{0,1\}}\left\{\mathsf{D}_1(y_1,y_0|s) - \mathsf{D}_2(y_1,y_0|s)\cdot d_2(P_{1-s}\|P_s)\right\}$$

where, for $p \geq 1$ and $s \in \{0,1\}$,

$$\mathsf{D}_p(y_1,y_0|s) \triangleq \left(\mathbb{E}\left[|y_1(X) - y_0(X)|^p \mid S = s\right]\right)^{1/p},$$

$\mathsf{D}_{\mathsf{TV}}(P_0\|P_1)$ is the total variation distance and $d_2(P_{1-s}\|P_s)$ is Marton's divergence. The ℓ_2-benefit-of-splitting $\epsilon_{split,2}$ can be upper bounded by

$$\min\left\{\min_{s\in\{0,1\}} \mathsf{D}_4(y_1, y_0|s)^2\cdot\sqrt{1-\mathsf{D}_{\mathsf{TV}}(P_0\|P_1)}, \frac{1}{4}\max_{s\in\{0,1\}}\mathsf{D}_2(y_1,y_0|s)^2\right\}$$

and lower bounded by

$$\max_{s\in\{0,1\}}\left\{\left(\frac{\mathsf{D}_1(y_1,y_0|s)}{\sqrt{\mathsf{D}_{\chi^2}(P_s\|P_{1-s})+1}+1}\right)^2\right\}$$

where $\mathsf{D}_{\chi^2}(P_s\|P_{1-s})$ is the chi-square divergence. The KL-benefit-of-splitting $\epsilon_{split,\mathsf{KL}}$ can be upper bounded by

$$\min\left\{2\mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - 2\mathsf{JS}(P_0\|P_1), \max_{s\in\{0,1\}}\mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(y_s(X)\|\tfrac{y_0(X)+y_1(X)}{2}\right)\mid S=s\right]\right\}$$

and lower bounded by

$$\mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}(P_0\|P_1)$$

where $\mathsf{JS}(\cdot\|\cdot)$ is the Jensen–Shannon divergence.

In particular, the proof of the lower bound of $\epsilon_{\text{split},2}$ relies on the following technical lemma.

Lemma 1. *For any measurable classifier $h : \mathcal{X} \to [0,1]$ and constant $0 \leq \epsilon < \mathsf{D}_2(y_1,y_0|0)^2$, if*

$$\mathbb{E}\left[(h(X) - y_0(X))^2 \mid S = 0\right] \leq \epsilon, \tag{4}$$

then $\mathbb{E}\left[(h(X)-y_1(X))^2 \mid S=1\right] \geq (A - B\sqrt{\epsilon})^2$*, where*

$$A \triangleq \mathsf{D}_1(y_1, y_0|1), \quad B \triangleq \sqrt{\mathsf{D}_{\chi^2}(P_1\|P_0)+1}.$$

Proof. Consider a convex optimization problem

$$\min_{h:\mathcal{X}\to[0,1]} \int (h(x)-y_1(x))^2 \mathrm{d}P_1(x),$$
$$\text{s.t.} \int (h(x)-y_0(x))^2 \mathrm{d}P_0(x) \leq \epsilon.$$

Computing the Gateaux derivative of the Lagrange multiplier gives the following optimal conditions [35, Theorem 6.6.1],

$$(h(x)-y_1(x))\mathrm{d}P_1(x) + \lambda(h(x)-y_0(x))\mathrm{d}P_0(x) = 0, \quad (5)$$
$$\lambda\left(\int (h(x)-y_0(x))^2 \mathrm{d}P_0(x) - \epsilon\right) = 0, \quad (6)$$
$$\lambda \geq 0, \quad (7)$$

which yields an optimal classifier:

$$h^*(x) = \frac{y_1(x)r(x) + \lambda y_0(x)}{r(x)+\lambda} \quad (8)$$

where $r(x) \triangleq \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}$. If $\lambda = 0$, then $h^*(x) = y_1(x)$ and

$$\mathbb{E}\left[(h^*(X)-y_0(X))^2 \mid S=0\right] = \mathsf{D}_2(y_1, y_0|0)^2.$$

However, this contradicts our assumptions that

$$\mathbb{E}\left[(h^*(X)-y_0(X))^2 \mid S=0\right] \leq \epsilon < \mathsf{D}_2(y_1, y_0|0)^2.$$

Hence, we have $\lambda > 0$. In this case, (6) and (8) imply

$$\int r(x)^2 \left(\frac{y_1(x)-y_0(x)}{r(x)+\lambda}\right)^2 \mathrm{d}P_0(x) = \epsilon. \quad (9)$$

By the expression of the optimal classifier in (8),

$$\begin{aligned}
&\mathbb{E}\left[(h^*(X)-y_1(X))^2 \mid S=1\right] \\
&\geq \left(\int \frac{\lambda|y_1(x)-y_0(x)|}{r(x)+\lambda}\mathrm{d}P_1(x)\right)^2 \\
&= \left(\mathsf{D}_1(y_1, y_0|1) - \int \frac{r(x)|y_1(x)-y_0(x)|}{r(x)+\lambda}\mathrm{d}P_1(x)\right)^2 \quad (10)
\end{aligned}$$

where the first step is due to the Cauchy-Schwarz inequality. By the Cauchy-Schwarz inequality again and (9), we have

$$\begin{aligned}
&\int \frac{r(x)|y_1(x)-y_0(x)|}{r(x)+\lambda}\mathrm{d}P_1(x) \\
&\leq \sqrt{\int r(x)^2 \left(\frac{y_1(x)-y_0(x)}{r(x)+\lambda}\right)^2 \mathrm{d}P_0(x) \int r(x)\mathrm{d}P_1(x)} \\
&= \sqrt{\epsilon\mathbb{E}\left[r(X) \mid S=1\right]} = \sqrt{\epsilon\left(\mathsf{D}_{\chi^2}(P_1\|P_0)+1\right)}. \quad (11)
\end{aligned}$$

Combining (10) and (11) leads to the desired conclusion. □

Now we consider extreme scenarios to verify the sharpness of the bounds in Theorem 1 and to understand when splitting classifiers benefits model performance the most (see Figure 1 for an illustration).

- Consider the setting where two groups share the same labeling function (i.e., $y_0 = y_1$). All the upper and lower bounds in Theorem 1 for the benefit-of-splitting under different loss functions become zero and, hence, the bounds are sharp. This is quite intuitive as one can use the labeling function y_0 (or y_1) as a group-blind classifier and it achieves perfect performance on both groups. Hence, there is no benefit of splitting classifiers.
- Consider the setting where two groups share the same unlabeled distribution (i.e., $P_0 = P_1$). The upper and lower bounds of $\epsilon_{\text{split},1}$ are both $\mathbb{E}\left[|y_1(X)-y_0(X)|\right]/2$, which is equal to $\epsilon_{\text{split},1}$. The bounds of $\epsilon_{\text{split},2}$ become

  $$\frac{1}{4}\mathbb{E}\left[|y_1(X)-y_0(X)|\right]^2 \leq \epsilon_{\text{split},2} \leq \frac{1}{4}\mathbb{E}\left[(y_1(X)-y_0(X))^2\right].$$

  If, in addition, $|y_0(x)-y_1(x)|$ is the same across all x, the upper and lower bounds become the same and, hence, are sharp. Finally, the bounds of $\epsilon_{\text{split},\mathsf{KL}}$ become

  $$\epsilon_{\text{split},\mathsf{KL}} \leq \max_{s\in\{0,1\}} \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(y_s(X)\|\tfrac{y_0(X)+y_1(X)}{2}\right)\right]$$
  $$\epsilon_{\text{split},\mathsf{KL}} \geq \mathbb{E}\left[\mathsf{JS}(y_0(X)\|y_1(X))\right].$$

  If, in addition, $\mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(y_0(X)\|(y_0(X)+y_1(X))/2\right)\right] = \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}\left(y_1(X)\|(y_0(X)+y_1(X))/2\right)\right]$, then the upper and lower bounds are equal. This extreme case indicates that when different groups have the same unlabeled distribution (i.e., $P_0 = P_1$), the benefit-of-splitting is determined by the disagreement between their labeling functions (i.e., large disagreement leads to high benefit).
- Consider the setting where two groups have unlabeled distributions lying on disjoint support sets. In this case, the upper bounds of $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split},\mathsf{KL}}$ are zero. In other words, there is no benefit of splitting classifiers when the unlabeled distributions are mutually singular. One can interpret this fact by considering a special group-blind classifier which mimics the labeling function of each group in the region where its unlabeled distribution lies. This classifier achieves perfect performance for each group. Note that such group-blind classifier exists since we do not restrict the space of potential classifiers and, hence, any (measurable) function could become a group-blind or split classifier.

To summarize, splitting classifiers brings the most benefit if two groups have similar unlabeled distributions and different labeling functions. This taxonomy of splitting appears for all the commonly used loss functions (i.e., ℓ_1, ℓ_2, and KL loss).

IV. An Efficient Procedure for Computing the Effect of Splitting

In the last section, we provide upper and lower bounds for the benefit-of-splitting under different loss functions. Here, we consider a different performance measure: false error rate. It turns out that the benefit-of-splitting under false error rate ($\epsilon_{\text{split},\mathsf{FER}}$ in Definition 3) has an equivalent expression which leads to an efficient procedure of computing $\epsilon_{\text{split},\mathsf{FER}}$.

Even with perfect knowledge of the underlying data distribution, computing the benefit-of-splitting directly from its definition is challenging. This is because the space of potential classifiers is unrestricted (i.e., any measurable function could

be used as a group-blind or split classifier) and solving optimization problems over this infinite-dimensional functional space could be intractable. One may attempt to circumvent this issue by restricting the classifiers over a hypothesis class. However, this has two limitations. First, it is unclear how to choose a hypothesis class in order to compute the benefit-of-splitting reliably. In fact, different hypothesis classes could result in completely different values of the benefit-of-splitting. Second, as evidenced in [36], training the optimal group-blind or split classifiers may suffer from a non-convexity issue.

We leverage the special form of the false error rate in Definition 3 and prove an equivalent expression of $\epsilon_{\text{split,FER}}$ below which can be computed by solving two small-scale convex programs. The objective functions of these convex programs have closed-form supergradients. Hence, they can be solved efficiently via standard solvers, such as (stochastic) mirror descent [37, 38]. The equivalent expression of $\epsilon_{\text{split,FER}}$ is given in the following theorem.

Theorem 2. *Assume that* $\Pr(Y = i, S = s) > 0$ *for any* $i, s \in \{0,1\}$. *The FER-benefit-of-splitting* $\epsilon_{\textit{split,FER}}$ *can be equivalently written as*

$$\max_{\boldsymbol{\mu}\in\Delta_4} \left\{ \sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X) \right)_- \right] \right\} - \max_{\substack{\boldsymbol{\nu}^{(s)}\in\Delta_2 \\ \textit{for } s\in\{0,1\}}} \left\{ \nu_1^{(s)} + \mathbb{E}\left[\left(\sum_{i\in\{0,1\}} \nu_i^{(s)}\phi_{s,i}(X) \right)_- \right] \right\}.$$

Here $\Delta_d \triangleq \{\boldsymbol{z} \in \mathbb{R}^d \mid \sum_{i=1}^d z_i = 1,\ z_i \geq 0\}$, $(a)_- \triangleq \min\{a, 0\}$, $\boldsymbol{\mu} \triangleq (\mu_{0,0}, \mu_{0,1}, \mu_{1,0}, \mu_{1,1})$, $\boldsymbol{\nu}^{(s)} \triangleq (\nu_0^{(s)}, \nu_1^{(s)})$, *and for* $s, i \in \{0,1\}$

$$\phi_{s,i}(x) \triangleq \frac{(1 - i - y_s(x)) \Pr(S = s \mid X = x)}{\Pr(Y = i, S = s)}.$$

Proof sketch. Recall that $\epsilon_{\text{split,FER}}$ can be written as

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) - \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h)$$

where the performance measure $L_s(h)$ is in Definition 3. The inf-max term can be equivalently written as

$$\inf_{h:\mathcal{X}\to[0,1]} \max_{\boldsymbol{\mu}\in\Delta_4} \Big\{ \sum_{s\in\{0,1\}} \mu_{s,0}\mathbb{E}\left[h(X) \mid Y = 0, S = s\right] + \mu_{s,1}\mathbb{E}\left[1 - h(X) \mid Y = 1, S = s\right] \Big\}. \tag{12}$$

The key step in our proof is to swap maximum and infimum in (12) by using Ky Fan's min-max theorem [21]. Then for a fixed $\boldsymbol{\mu}$, the optimal classifier owns a closed-form expression. After some algebraic manipulations, (12) becomes the first convex program in the equivalent expression of $\epsilon_{\text{split,FER}}$. In the same vein, the another term $\max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h)$ becomes the second convex program. □

Next, we show that the objective functions of the convex programs in Theorem 2 have closed-form supergradients.

Proposition 1. *Under the setting in Theorem 2, the objective functions* $g : \Delta_4 \to \mathbb{R}$ *and* $g_s : \Delta_2 \to \mathbb{R}$

$$g(\boldsymbol{\mu}) \triangleq \sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X) \right)_- \right]$$

$$g_s(\boldsymbol{\nu}) \triangleq \nu_1 + \mathbb{E}\left[\left(\sum_{i\in\{0,1\}} \nu_i\phi_{s,i}(X) \right)_- \right]$$

have a closed-form supergradient, respectively:

$$\left(i + \mathbb{E}\left[\psi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{s',i'} \mu_{s',i'}\phi_{s',i'}(X) < 0 \Big] \Big| S = s \right] \right)_{s,i}$$

$$\left(i + \mathbb{E}\left[\psi_{s,i}(X) \cdot \mathbb{I}\Big[\sum_{i'} \nu_{i'}\phi_{s,i'}(X) < 0 \Big] \Big| S = s \right] \right)_{i}$$

where $\mathbb{I}[\cdot]$ *is the indicator function and*

$$\psi_{s,i}(x) \triangleq \frac{1 - i - y_s(x)}{\Pr(Y = i \mid S = s)}, \quad s, i \in \{0,1\}.$$

When the underlying data distribution is known, one can compute $\epsilon_{\text{split,FER}}$ by solving the convex programs in Theorem 2 via standard tools, such as mirror descent, with convergence guarantees [38]. This is non-trivial because, as stated before, computing $\epsilon_{\text{split,FER}}$ directly from its definition could be intractable.

In practice, when the underlying data distribution is unknown, one can first approximate the conditional distribution $\Pr(S = 1|X = x)$ and the labeling functions $y_0(x)$, $y_1(x)$ by training three well-calibrated binary classifiers. These classifiers will be called when computing the supergradient of the objective functions (Proposition 1). Our procedure can be understood through the following two steps:

1) training a classifier to identify the sensitive attribute using input features and a classifier for each group to predict label using input features;
2) solving (convex) programs with these classifiers in hand.

We remark that this two-step approach has also appeared in [e.g., 12, 39] for designing "fair" classifiers. Numerical results are available in the full version of this paper [1].

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under grants CIF 1900750, CAREER 1845852, and IIS 1926925 and an Amazon Research Award. The work of M. Diaz was supported in part by the Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) under grant IA101021.

REFERENCES

[1] H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "To split or not to split: The impact of disparate treatment in classification," *arXiv preprint arXiv:2002.04788*, 2020.

[2] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[3] EEOC, "Uniform guidelines on employee selection procedures," 1979. [Online]. Available: https://www.eeoc.gov/policy/docs/qanda_clarify_procedures.html

[4] Federal Trade Commission (FTC), "Equal Credit Opportunity Act (ECOA)," 2020. [Online]. Available: https://www.fdic.gov/resources/supervision-and-examinations/consumer-compliance-examination-manual/documents/5/v-7-1.pdf

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 259–268.

[6] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[7] T. L. Beauchamp and J. F. Childress, *Principles of biomedical ethics*. Oxford University Press, USA, 2001.

[8] B. Ustun, Y. Liu, and D. Parkes, "Fairness without harm: Decoupled classifiers with preference guarantees," in *Proc. 36th International Conference on Machine Learning*, 2019, pp. 6373–6382.

[9] N. Martinez, M. Bertran, and G. Sapiro, "Fairness with minimal harm: A pareto-optimal approach for healthcare," *arXiv preprint arXiv:1911.06935*, 2019.

[10] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller, "From parity to preference-based notions of fairness in classification," in *Advances in Neural Information Processing Systems*, 2017, pp. 229–239.

[11] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled classifiers for group-fair and efficient machine learning," in *Proc. 1st Conference on Fairness, Accountability and Transparency*, 2018, pp. 119–133.

[12] H. Wang, B. Ustun, and F. P. Calmon, "Repairing without retraining: Avoiding disparate impact with counterfactual distributions," in *Proc. 36th International Conference on Machine Learning*, 2019, pp. 6618–6627.

[13] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan, "Algorithmic fairness," in *Aea papers and proceedings*, vol. 108, 2018, pp. 22–27.

[14] A. Blum and K. Stangl, "Recovering from biased data: Can fairness constraints improve accuracy?" *arXiv preprint arXiv:1912.01094*, 2019.

[15] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, 2019.

[16] H. H. Zhou, Y. Zhang, V. K. Ithapu, S. C. Johnson, G. Wahba, and V. Singh, "When can multi-site datasets be pooled for regression? hypothesis tests, l_2-consistency and neuroscience applications," in *Proc. 34th International Conference on Machine Learning*, 2017, pp. 4170–4179.

[17] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, vol. 79, no. 1-2, pp. 151–175, 2010.

[18] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *22nd Conference on Learning Theory*, 2009.

[19] A. G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity," *Continuing Education in Anaesthesia Critical Care & Pain*, vol. 8, no. 6, pp. 221–223, 2008.

[20] L. D. Brown and M. G. Low, "A constrained risk inequality with applications to nonparametric functional estimation," *The Annals of Statistics*, vol. 24, no. 6, pp. 2524–2535, 1996.

[21] K. Fan, "Minimax theorems," *Proc. National Academy of Sciences of the United States of America*, vol. 39, no. 1, p. 42, 1953.

[22] A. B. Tsybakov, *Introduction to nonparametric estimation*. Springer Science & Business Media, 2008.

[23] Y. Wu and P. Yang, "Minimax rates of entropy estimation on large alphabets via best polynomial approximation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3702–3720, 2016.

[24] J. Jiao, Y. Han, and T. Weissman, "Minimax estimation of the l_1 distance," *IEEE Trans. Inf. Theory*, vol. 64, no. 10, pp. 6672–6706, 2018.

[25] Y. Polyanskiy and Y. Wu, "Dualizing Le Cam's method, with applications to estimating the unseens," *arXiv preprint arXiv:1902.05616*, 2019.

[26] A. Xu and M. Raginsky, "Information-theoretic lower bounds on bayes risk in decentralized estimation," *IEEE Trans. Inf. Theory*, vol. 63, no. 3, pp. 1580–1600, 2016.

[27] J. Duchi, M. J. Wainwright, and M. I. Jordan, "Local privacy and minimax bounds: Sharp rates for probability estimation," in *Advances in Neural Information Processing Systems*, 2013, pp. 1529–1537.

[28] S. B. David, T. Lu, T. Luu, and D. Pal, "Impossibility theorems for domain adaptation," in *Proc. Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 129–136.

[29] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observation," *studia scientiarum Mathematicarum Hungarica*, vol. 2, pp. 229–318, 1967.

[30] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," in *Advances in Neural Information Processing Systems*, 2017, pp. 5680–5689.

[31] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[32] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the (im) possibility of fairness," *arXiv preprint arXiv:1609.07236*, 2016.

[33] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[34] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations," *arXiv preprint arXiv:1906.08386*, 2019.

[35] A. J. Kurdila and M. Zabarankin, *Convex functional analysis*. Springer Science & Business Media, 2006.

[36] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, 2017, pp. 962–970.

[37] A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization.* Wiley, 1983.

[38] A. Beck and M. Teboulle, "Mirror descent and nonlinear projected subgradient methods for convex optimization," *Operations Research Letters*, vol. 31, no. 3, pp. 167–175, 2003.

[39] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Proc. 1st Conference on Fairness, Accountability and Transparency*, 2018, pp. 107–118.