To Split or Not to Split: The Impact of Disparate Treatment in Classification

Hao Wang, *Student Member, IEEE,* Hsiang Hsu, *Student Member, IEEE,* Mario Diaz, *Member, IEEE,* and Flavio P. Calmon, *Member, IEEE*

***Abstract*—Disparate treatment occurs when a machine learning model produces different decisions for individuals based on a legally protected or 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). In the information-theoretic regime, we introduce the benefit-of-splitting for quantifying the performance improvement by splitting classifiers. Computing the benefit-of-splitting directly from its definition could be intractable since it 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. In the finite sample regime, splitting is not necessarily beneficial and we provide data-dependent bounds to understand this effect. Finally, we validate our theoretical results through numerical experiments on both synthetic and real-world datasets.**

***Index Terms*—Trustworthy machine learning, fairness, domain adaptation, f-divergence, converse bounds.**

I. Introduction

A machine learning (ML) model exhibits *disparate treatment* [1] if it treats similar data points from distinct individuals differently based on a sensitive attribute (e.g., age, sex). In applications such as hiring, the existence of disparate treatment can be illegal [2]. 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 [3]–[5]. 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 [6]).

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 [7] and equalized odds [8]. 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") [9] become more appropriate moral principles for fairness [10]–[12]. 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 [10], [13]–[15]. This probability distribution can vary significantly conditioned on a sensitive attribute due to, for example, inherent differences between groups [13], differences in labeling [16], and differences in sampling [17]. 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 [18]: 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:

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

First, we show that in the information-theoretic regime where 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

Manuscript received July 11, 2020; accepted March 30, 2021. 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 PAPIIT grant IA101021. Some parts of this paper were presented at the 2020 Symposium on Foundations of Responsible Computing (FORC).

H. Wang, H. Hsiang, and F. P. Calmon are with Harvard University, Cambridge, MA 02138 USA (e-mail:{hao_wang,hsianghsu}@g.harvard.edu; flavio@seas.harvard.edu).
M. Diaz is with the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS), Universidad Nacional Autónoma de México, Mexico City, Mexico (e-mail: mario.diaz@sigma.iimas.unam.mx).

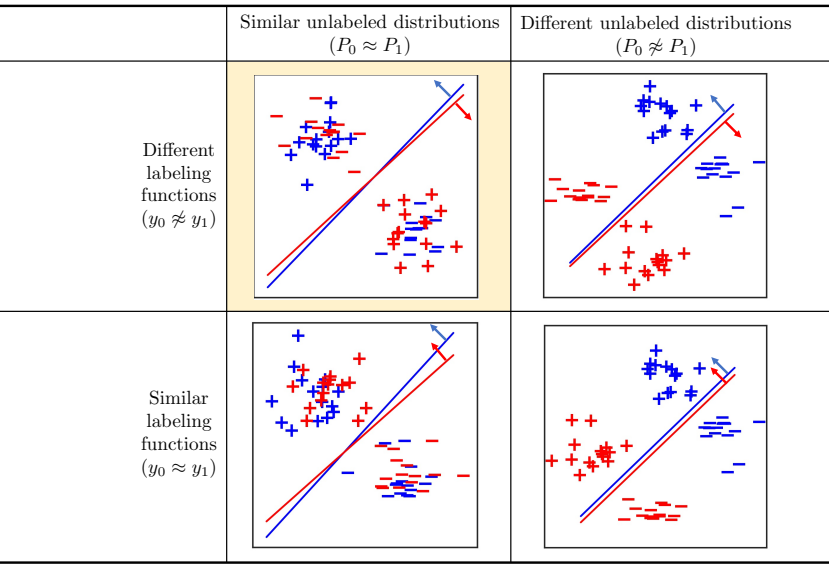

Fig. 1: The taxonomy of splitting based on two different factors. Samples from two groups are depicted in red and blue, respectively, and their labels are represented by $+$, $-$. Each group's labeling function is shown with the corresponding color and the arrows indicate the regions where the points 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).

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 and cross entropy 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]. Based on these two factors, our upper bounds in Theorem 1 indicate that splitting does not produce much benefit if the labeling functions are similar or the unlabeled distributions are different; our lower bounds in Theorem 1 indicate that splitting benefits the most if two groups' labeling functions are different and unlabeled distributions are similar. Furthermore, our lower bounds in Theorem 1 lead to an impossibility (i.e., converse) result for group-blind classifiers: under certain precise conditions, using a group-blind classifier will always suffer from an inherent accuracy trade-off between different groups and splitting classifiers can reconcile this issue. This converse result is information-theoretic: a data scientist cannot overcome this limit by using more samples or altering the hypothesis class.

In Section IV, we consider false error rate as a performance measure since in applications such as medical diagnostics, high false error rate could result in unintentional harm [21]. 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 2). Combining these two results leads to an efficient procedure (Algorithm 1) for computing the benefit-of-splitting. We validate our procedure through numerical experiments on synthetic datasets in Section VI-A. When the underlying data distribution is known, our procedure has a provable convergence guarantee and returns the precise values of the benefit-of-splitting. When the underlying data distribution is unknown, our procedure may suffer from approximation errors but still outperforms more naive empirical approaches.

The aforementioned results capture the benefit-of-splitting from an information-theoretic perspective where the underlying data distributions are assumed to be known and the space of potential classifiers is unrestricted. In Section V, we consider the effect of splitting classifiers in a more practical setting where group-blind and split classifiers are restricted over the same hypothesis class (e.g., logistic regressions) and the underlying distribution is accessed only through finitely many i.i.d. samples. In this case, splitting classifiers is not necessarily beneficial since the group with less samples may suffer from overfitting. To quantify the effect of splitting classifiers, we analyze the *sample-limited benefit-of-splitting*. We derive upper and lower bounds for the benefit-of-splitting in this regime in Theorem 4. These bounds disentangle three factors which determine the effect of splitting classifiers in practice: (i) disagreement between optimal (split) classifiers and training error associated with these optimal classifiers; (ii) similarity between (empirical) unlabeled distributions; and (iii) model complexity and number of samples. The first two factors are analogous to the ones that affect the benefit-of-splitting in the information-theoretic regime: when the hypothesis class is complex enough and the sample size tends to infinity, the optimal classifiers approximate the labeling functions and the empirical unlabeled distributions converge to the true unlabeled distributions. Finally, we illustrate how these factors determine the performance impact of splitting classifiers through experiments on 40 datasets downloaded from OpenML [22].

The proof techniques of this paper are based on fundamental tools found in statistics, such as Brown-Low's two-points lower bound [23], and methods in convex analysis, such as Ky Fan's min-max theorem [24]. These tools are widely used in applications such as non-parametric estimation [25], and are useful for analyzing the min-max risk in statistical settings [26]–[30]. Furthermore, the factors that we provide for understanding the effect of splitting classifiers are inspired by

[1] We borrow the terms "labeling function" and "unlabeled distribution" from the domain adaptation literature [19], [20]. The labeling function takes a data point as an input and produces a probability of its binary label being 1 and the unlabeled distribution is a (marginal) probability distribution of the unlabeled data. 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-B.

the necessary and sufficient conditions of domain adaptation learnability in Ben-David *et al.* [31].

The rest of this paper is organized as follows. In the remainder of this section, we review related works and present notation adopted in this paper. In Section II, we introduce the main object of interest: the benefit-of-splitting. Under different performance measures, we provide upper and lower bounds for the benefit-of-splitting in Section III and present an efficient procedure for computing the benefit-of-splitting in Section IV. The effect of splitting classifiers in the finite sample regime is studied in Section V. Finally, we illustrate our results through numerical experiments in Section VI and provide conclusion remarks and future works in Section VII.

A. Related Work

a) Privacy: Fairness and privacy are closely connected and central to trustworthy machine learning. In this paper, we study the impact of disparate treatment from an information-theoretic perspective: we assume the underlying data distribution is known and analyze how the different distributions between groups affect the performance improvement by splitting classifiers. In this regard, our present work relates with studies on information-theoretic privacy, see e.g., [32]–[38]. These efforts explore the fundamental limits of privacy-utility trade-offs by also assuming the underlying data distribution is known. For example, Makhdoumi *et al.* [39] introduce privacy funnel method for solving the privacy-utility trade-offs and connect it with the information bottleneck [40], and this connection is further studied in [41]. Kairouz *et al.* [42] study the trade-offs between local differential privacy [30], [43], [44] and utility functions measured by f-divergence [45]. Besides analyzing the fundamental limits, there are works [46]–[49] on designing privacy mechanisms which enable a certain level of utility to be obtained from the disclosed datasets while controlling private information leakage. The robustness of the privacy mechanisms is analyzed in [50], [51] when these privacy mechanisms are constructed by using finitely many samples. We follow a similar line of analysis in order to understand the effect of splitting classifiers in the finite sample regime and complement our bounds for the benefit-of-splitting by incorporating additional factors such as sample size and model complexity.

b) Domain adaptation: A standard assumption in ML is that the training and testing data are drawn from the same underlying probability distribution. Domain adaptation [19], [20], [52] and transfer learning [53], [54] consider a more general setting where models are trained on a source domain and deployed on a (different) target domain. A common assumption therein is known as *covariate shift*, which requires the source and target domain share the same labeling function. In this paper, we prove (see Theorem 1) that if the covariate shift assumption is violated and two groups' unlabeled distributions are similar, then no classifier can perform well on both groups. In this regard, our work is connected to Ben-David *et al.* [31] which present impossibility results on domain adaptation learnability. Compared to [31], Theorem 1 characterizes an information-theoretic fundamental limit which cannot be circumvented by using a large number of samples or a carefully designed hypothesis class. Furthermore, the lower bound in Theorem 1 serves as a complementary statement to the upper bounds in domain adaptation [19], [20]. These bounds jointly describe the range of the loss a data scientist may incur by training a model on the source domain and deploying on the target domain.

c) Fair ML: ML models have been increasingly used in applications of individual-level consequences, ranging from recidivism prediction [55] and lending [56] to healthcare [57]. A number of works in fair ML aim at understanding why discrimination happens [58]–[66]; how it can be quantified [67]–[69]; and how it can be reduced [70]–[79]. There are also an increasing number of studies that take causality into account for understanding and mitigating discrimination [80]–[84]. We build on a line of recent results on decoupling predictive models for improving accuracy-fairness trade-offs see e.g., [10], [12], [13], [85], [86]. For example, Ustun *et al.* [10] introduce a tree structure to recursively choose sensitive attributes for decoupling. Lipton *et al.* [85] show that using group-blind classifiers could be suboptimal in terms of trading off accuracy and fairness. The work closest to ours is Dwork *et al.* [13] which present a decoupling technique to learn separate models for different groups. A detailed comparison with [13] is given in Section V-B.

B. 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[2] 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 (i.e., P_0 and P_1), we recall Csiszár's f-divergence [45]. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$ and P, Q be two probability distributions over $\mathcal{X}$. The f-divergence between P and Q is defined by

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

Some examples of f-divergence are included in Appendix A.

The proofs of some of our main results (Lemma 2 and Theorem 2) rely on Ky Fan's min-max theorem [24]. As a

[2]For the sake of illustration, we assume that the sensitive attribute S is binary but our results can be extended to a setting of multi-groups. Furthermore, split classifiers can be applied to a scenario where multiple subgroups overlap [74], [87] since individuals belonging to both groups can opt for either one of the split classifiers.

reminder, a function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is said to be concave-like on $\mathcal{X}$ if, for any two elements $x_1, x_2 \in \mathcal{X}$ and $\lambda \in [0,1]$, there exists an element $x_0 \in \mathcal{X}$ such that for all $y \in \mathcal{Y}$

$$f(x_0, y) \geq \lambda f(x_1, y) + (1-\lambda) f(x_2, y).$$

Similarly, f is said to be convex-like on $\mathcal{Y}$, if for any two elements $y_1, y_2 \in \mathcal{Y}$ and $\lambda \in [0,1]$, there exists an element $y_0 \in \mathcal{Y}$ such that for all $x \in \mathcal{X}$

$$f(x, y_0) \leq \lambda f(x, y_1) + (1-\lambda) f(x, y_2).$$

A function $g : \mathcal{X} \to \mathbb{R}$ is called upper semicontinuous on a metric space $\mathcal{X}$ if for every point $x_0 \in \mathcal{X}$, $\limsup_{x \to x_0} g(x) \leq g(x_0)$. Next, we recall[3] Ky Fan's min-max theorem [24].

Lemma 1 ([24, Theorem 2])**.** *Let $\mathcal{X}$ be a compact Hausdorff space and $\mathcal{Y}$ an arbitrary set (not topologized). Let f be a real-valued function on $\mathcal{X} \times \mathcal{Y}$ such that, for every $y \in \mathcal{Y}$, $f(\cdot, y)$ is upper semicontinuous on $\mathcal{X}$. If f is concave-like on $\mathcal{X}$ and convex-like on $\mathcal{Y}$, then*

$$\inf_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} f(x,y) = \max_{x \in \mathcal{X}} \inf_{y \in \mathcal{Y}} f(x,y). \tag{2}$$

II. The Benefit-of-Splitting

We study the impact of disparate treatment by comparing the performance between optimal group-blind and split classifiers. Recall that a ML model exhibits disparate treatment if it explicitly uses a sensitive attribute to produce an output. 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 (across all measurable functions from $\mathcal{X}$ to $[0,1]$) 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 each $s \in \{0,1\}$, let $P_{X,Y|S=s}$ be a fixed probability distribution and $L_s(\cdot)$ be a performance measure, we define the benefit-of-splitting as

$$\epsilon_{\text{split}} \triangleq \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} L_s(h) - \max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} L_s(h), \tag{3}$$

[3]We apply Ky Fan's min-max theorem to the function $-f$ instead of f.

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

$$\epsilon_{\text{split}} = \max_{s\in\{0,1\}} L_s(h^*) - \max_{s\in\{0,1\}} L_s(h_s^*). \tag{4}$$

In practice, a data scientist may restrict the type of classifiers by fixing a hypothesis class (e.g., logistic regressions). The benefit-of-splitting can be adapted for capturing the effect of splitting classifiers in this case (see Definition 5).

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 have been widely used in fair ML see e.g., [13] and domain adaptation see e.g., [19]. These loss functions 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 $p, q \in [0,1]$.

Remark 1. Another widely used loss function is cross entropy $L_s(h) = \mathbb{E}\left[\mathsf{H}(y_s(X), h(X)) \mid S = s\right]$ where for $p, q \in [0,1]$, $\mathsf{H}(p,q) \triangleq -p\log q - (1-p)\log(1-q)$. Since $\mathsf{H}(p,q) = \mathsf{D}_{\mathsf{KL}}(p\|q) + \mathsf{H}(p)$, the analysis of the benefit-of-splitting under cross entropy is essentially the same as the analysis of $\epsilon_{\text{split},\mathsf{KL}}$ (see Appendix B-C).

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[4] between (generalized) false positive rate and (generalized) false negative rate [67]. In healthcare, assuring low false error rate is as important as guaranteeing high accuracy since patients could suffer from harm due to a classifier's false error rate [21]. For example, the false negative diagnosis may delay treatment in patients who are critically ill; the false positive diagnosis could lead to an unnecessary treatment. Furthermore, a classifier with high accuracy does not necessarily mean it has low false error rate. Hence, we consider how split classifiers reduce the false error rate by recasting the benefit-of-splitting under this performance measure.

Definition 3. The FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ 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, discussed by Hardt *et al.* [8], is a commonly used group fairness measure that requires different groups to have (approximately) the same false positive and false negative rates. Specifically, a probabilistic classifier $h:\mathcal{X}\to[0,1]$ satisfies equalized odds [8], [67] if

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

Under this definition, classifiers are considered "unfair" if their false positive rate or false negative rate vary across different groups. However, imposing equalized odds constraints may lead to a significant performance reduction in classification [60], [88]–[90]. In contrast, the benefit-of-splitting definition studied in this work aims to capture the principles of non-maleficence and beneficence [9]: 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 (see 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.

Before stating the main result (i.e., bounds for the benefit-of-splitting), we prove a lemma first which converts the definition of the benefit-of-splitting into a single variable optimization problem. This lemma will be used in the proof of our lower bounds.

Lemma 2. *The benefit-of-splitting under different loss functions in Definition 2 have equivalent expressions*

$$\epsilon_{split,1}=\sup_{\omega\in[0,1]}(1-\omega)\int_{\mathcal{A}_\omega}|y_1(x)-y_0(x)|dP_1(x)+\omega\int_{\mathcal{A}_\omega^c}|y_1(x)-y_0(x)|dP_0(x),$$
$$\epsilon_{split,2}=\sup_{\omega\in[0,1]}\omega(1-\omega)\int\frac{(y_1(x)-y_0(x))^2dP_0(x)dP_1(x)}{\omega dP_0(x)+(1-\omega)dP_1(x)},$$
$$\epsilon_{split,\mathsf{KL}}=\sup_{\omega\in[0,1]}\mathsf{JS}_\omega(P_{X,Y|S=0}\|P_{X,Y|S=1})-\mathsf{JS}_\omega(P_0\|P_1),$$

where $\mathcal{A}_\omega\triangleq\left\{x\in\mathcal{X}\mid\frac{dP_0(x)}{dP_1(x)}\geq\frac{1-\omega}{\omega}\right\}$ *and* $\mathsf{JS}_\omega(\cdot\|\cdot)$ *is the Jensen-Shannon divergence.*

Proof. See Appendix B-A. □

We provide upper and lower bounds for $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split,KL}}$, respectively, in Theorem 1. These bounds rely on two main factors: (i) disagreement between different groups' labeling functions and (ii) similarity between their unlabeled distributions. In particular, the second factor is captured by a certain f-divergence [45] (see Appendix A for some examples of f-divergence).

Now we consider extreme scenarios to verify the sharpness of the bounds 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 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 $\epsilon_{\text{split,KL}}$ become

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

  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

[4]Our analysis can be extended to any convex combination of false positive rate and false negative rate.

Theorem 1. *The ℓ_1-benefit-of-splitting can be upper and lower bounded*

$$\epsilon_{\textit{split},1} \leq \min\left\{ \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]} \cdot \sqrt{1-\mathsf{D_{TV}}(P_0\|P_1)}, \right.$$
$$\left. \frac{1}{2}\max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=s\right] \right\},$$
$$\epsilon_{\textit{split},1} \geq \frac{1}{2}\max_{s\in\{0,1\}} \left\{ \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=s\right] - \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right]} \cdot d_2(P_{1-s}\|P_s) \right\},$$

where $\mathsf{D_{TV}}(P_0\|P_1)$ is the total variation distance and $d_2(P_{1-s}\|P_s)$ is Marton's divergence. The ℓ_2-benefit-of-splitting can be upper and lower bounded

$$\epsilon_{\textit{split},2} \leq \min\left\{ \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X)-y_0(X))^4 \mid S=s\right]} \cdot \sqrt{1-\mathsf{D_{TV}}(P_0\|P_1)}, \right.$$
$$\left. \frac{1}{4}\max_{s\in\{0,1\}} \mathbb{E}\left[(y_1(X)-y_0(X))^2 \mid S=s\right] \right\},$$
$$\epsilon_{\textit{split},2} \geq \max_{s\in\{0,1\}} \left\{ \left(\frac{\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=s\right]}{\sqrt{\mathsf{D}_{\chi^2}(P_s\|P_{1-s})+1}+1} \right)^2 \right\},$$

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

$$\epsilon_{\textit{split},\mathsf{KL}} \leq \min\left\{ 2\mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - 2\mathsf{JS}(P_0\|P_1),\ \max_{s\in\{0,1\}} \mathbb{E}\left[\mathsf{D_{KL}}\left(y_s(X) \| \frac{y_0(X)+y_1(X)}{2} \right) \mid S=s \right] \right\},$$
$$\epsilon_{\textit{split},\mathsf{KL}} \geq \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}(P_0\|P_1),$$

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

Proof. See Appendix B-B. □

---

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, $\mathsf{D_{TV}}(P_0\|P_1)=1$ and $\mathsf{JS}(P_0\|P_1)=\log 2$. Hence, the upper bounds of $\epsilon_{\text{split},1}$ and $\epsilon_{\text{split},2}$ become zero. Furthermore,

$$0 \leq \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}(P_0\|P_1)$$
$$= \mathsf{JS}(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \log 2 \leq 0.$$

  where the last step is because the Jensen–Shannon divergence is always upper bounded by $\log 2$. Therefore, the upper bound of $\epsilon_{\text{split},\mathsf{KL}}$ is zero as well. 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 a group-blind classifier exists since we do not restrict the space of potential classifiers and, hence, any (measurable) function could become a classifier.

To summarize, from an information-theoretic perspective, splitting classifiers benefits the most 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).

Recall that the benefit-of-splitting (see Definition 1) measures the performance improvement by using the optimal split classifiers compared to deploying the optimal group-blind classifier across all groups. Here model performance is quantified in terms of the disadvantaged group (i.e., the group with the worst performance). We end this section by considering the Bayes risk as an alternative way of measuring model performance[5]. Specifically, the performance of a group-blind classifier h and a set of split classifiers $\{h_s\}_{s\in\{0,1\}}$, respectively, is measured by

$$\sum_{s\in\{0,1\}} \Pr(S=s)\cdot\mathbb{E}\left[|h(X)-y_s(X)| \mid S=s\right],$$
$$\sum_{s\in\{0,1\}} \Pr(S=s)\cdot\mathbb{E}\left[|h_s(X)-y_s(X)| \mid S=s\right].$$

They can be equivalently written as

$$\mathbb{E}\left[|h(X)-y_S(X)|\right] \quad \text{and} \quad \mathbb{E}\left[|h_S(X)-y_S(X)|\right].$$

The performance difference between the optimal group-blind and split classifiers leads to the following definition.

Definition 4. We define the population-benefit-of-splitting as

$$\epsilon_{\text{split,pop}} \triangleq \inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[|h(X)-y_S(X)|\right]$$
$$- \inf_{\substack{h_s:\mathcal{X}\to[0,1] \\ \text{for } s\in\{0,1\}}} \mathbb{E}\left[|h_S(X)-y_S(X)|\right].$$

[5]For the sake of illustration, in what follows we only consider the ℓ_1 loss.

The population-benefit-of-splitting is upper bounded by the benefit-of-splitting (i.e., $\epsilon_{\text{split,pop}} \leq \epsilon_{\text{split},1}$) since the Bayes risk is upper bounded by the worst-case risk and the split classifiers $\{y_s\}_{s\in\{0,1\}}$ composed by the labeling functions can achieve zero risk. Hence, the upper bound of $\epsilon_{\text{split},1}$ in Theorem 1 naturally translates into an upper bound of $\epsilon_{\text{split,pop}}$. Next, we provide alternative bounds for $\epsilon_{\text{split,pop}}$ which reveal an additional factor influencing $\epsilon_{\text{split,pop}}$.

Proposition 1. *Assume* $\Pr(S=0) \leq 0.5$. *The population-benefit-of-splitting can be upper and lower bounded:*

$$\epsilon_{split,pop} \leq \Pr(S=0)\cdot \mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right],$$
$$\epsilon_{split,pop} \geq \Pr(S=0)\Big(\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] - E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)\Big),$$

where $E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)$ *is the* E_γ*-divergence with* $\gamma = \Pr(S=1)/\Pr(S=0)$.

Proof. See Appendix B-D. □

Remark 2. The E_γ-divergence plays an important role in Bayesian statistical hypothesis testing [91], [92]. Since $\gamma \to E_\gamma(P\|Q)$ is non-increasing and $E_1(P\|Q) = \mathsf{D}_{\mathsf{TV}}(P\|Q)$ [92], we can further lower bound $\epsilon_{\text{split,pop}}$ by using the total variation distance

$$\epsilon_{\text{split,pop}} \geq \Pr(S=0)(\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1)).$$

The E_γ-divergence relates with the DeGroot statistical information [93] through (see Equation (421) in [91])

$$\mathcal{I}_p(P\|Q) = \begin{cases} pE_{\frac{1-p}{p}}(P\|Q) & p\in(0,\frac{1}{2}] \\ (1-p)E_{\frac{p}{1-p}}(Q\|P) & p\in[\frac{1}{2},1). \end{cases}$$

Hence, we can write our lower bound of $\epsilon_{\text{split,pop}}$ equivalently as

$$\Pr(S=0)\cdot\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right] - \mathcal{I}_{\Pr(S=0)}(P_0\|P_1).$$

As shown in Proposition 1, the population-benefit-of-splitting is affected not only by the above-mentioned two factors (i.e., disagreement between labeling functions and similarity between unlabeled distributions) but also by the percentage of the minority group over the whole population. This reveals a caveat of the population-benefit-of-splitting: the minority group can be underrepresented when one designs a group-blind classifier by minimizing the loss over the whole population. In contrast, the benefit-of-splitting (see Definition 2) does not rely on the probability of the sensitive attribute and, hence, represents each group equally.

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 kinds of loss functions. Here, we consider a different performance measure: false error rate. It turns out that the benefit-of-splitting under false error rate, denoted by $\epsilon_{\text{split},\mathsf{FER}}$ (see Definition 3), has an equivalent expression which leads to an efficient procedure of computing $\epsilon_{\text{split},\mathsf{FER}}$.

Even with the 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 group-blind or split classifiers) 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 naive approach has two limitations. First, it is unclear how to choose a hypothesis class in order to compute the benefit-of-splitting reliably. We will show in Example 1 that different hypothesis classes could result in completely different values of the benefit-of-splitting. Second, as evidenced in [94], 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 $\epsilon_{\text{split},\mathsf{FER}}$ 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 [95], [96]. When the data distribution is known, our procedure returns the precise values of $\epsilon_{\text{split},\mathsf{FER}}$ without the need of training optimal group-blind and split classifiers. The equivalent expression of $\epsilon_{\text{split},\mathsf{FER}}$ is given in the following theorem.

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

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

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

$$\phi_{s,i}(x) \triangleq \frac{(1-i-y_s(x))\Pr(S=s\mid X=x)}{\Pr(Y=i,S=s)}. \tag{5}$$

Proof. See Appendix C-A. □

Remark 3. We demonstrate a proof sketch of Theorem 2. Note $\epsilon_{\text{split},\mathsf{FER}}$ is composed by $\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}L_s(h)$ and $\max_{s\in\{0,1\}}\inf_{h:\mathcal{X}\to[0,1]}L_s(h)$. The first term can be equivalent written as

$$\inf_{h:\mathcal{X}\to[0,1]}\max_{\boldsymbol{\mu}\in\Delta_4}\Big\{\sum_{s\in\{0,1\}}\mu_{s,0}\mathbb{E}\left[h(X)\mid Y=0,S=s\right] + \mu_{s,1}\mathbb{E}\left[1-h(X)\mid Y=1,S=s\right]\Big\}. \tag{6}$$

The key step in our proof is to swap maximum and infimum in (6) by using Ky Fan's min-max theorem [24] (see

Lemma 1). Then for a fixed $\boldsymbol{\mu}$, the optimal classifier owns a closed-form expression. After some algebraic manipulations, (6) 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 2. *Under the same notations and assumptions in Theorem 2, the functions $g:\Delta_4\to\mathbb{R}$ and $g_s:\Delta_2\to\mathbb{R}$ with $s\in\{0,1\}$ defined as*

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

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

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

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

Proof. See Appendix C-B. □

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 [96]. 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 (see Proposition 2). We summarize our procedure of computing $\epsilon_{\text{split,FER}}$ in Algorithm 1 where stochastic mirror descent is used for solving the convex programs in Theorem 2. The numerical results are deferred to Section VI-A.

Our procedure can be understood through the following two steps:

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

We remark that this two-step approach has also appeared in e.g., [14], [72] for designing "fair" classifiers.

Algorithm 1 Compute $\epsilon_{\text{split,FER}}$ via stochastic mirror descent.

Input:
dataset: $\mathcal{D}=\{(x_i,y_i,s_i)\}_{i=1}^n$, max number of iterations: T, step size: $\{\eta_t\}_{t=1}^T$
Initialize
$\mathcal{I}_0 \leftarrow \{i=1,\cdots,n \mid s_i=0\}$ ▷ indices of points with $s_i=0$
$\mathcal{D}_0 \leftarrow (x_i,y_i)$ for $i\in\mathcal{I}_0$ ▷ points with $s_i=0$
$\mathcal{D}_1 \leftarrow (x_i,y_i)$ for $i\notin\mathcal{I}_0$ ▷ points with $s_i=1$
approximate $\Pr(S=1\mid X=x)$
approximate $y_0(x)$ and $y_1(x)$
$\boldsymbol{\mu}\leftarrow(0.25,0.25,0.25,0.25)$ and $\boldsymbol{\nu}^{(s)}\leftarrow(0.5,0.5)$
for t = 1,2, $\cdots$,T **do**
  draw unlabeled sample $x_{0,t}$, $x_{1,t}$ from $\mathcal{D}_0$, $\mathcal{D}_1$
  pick $\boldsymbol{w}\in\partial g(\boldsymbol{\mu})$ and $\boldsymbol{w}^{(s)}\in\partial g_s(\boldsymbol{\nu}^{(s)})$
  $\mu_j \leftarrow \mu_j\exp(\eta_t w_j)/\sum_{j'}\mu_{j'}\exp(\eta_t w_{j'})$
  $\nu_j^{(s)} \leftarrow \nu_j^{(s)}\exp(\eta_t w_j^{(s)})/\sum_{j'}\nu_{j'}^{(s)}\exp(\eta_t w_{j'}^{(s)})$
end for
return: $g(\boldsymbol{\mu}) - \max_{s\in\{0,1\}} g_s(\boldsymbol{\nu}^{(s)})$

V. Splitting in Practice

So far we have studied the benefit-of-splitting from an information-theoretic view as we assume the underlying data distribution is known and do not restrict the space of potential classifiers. In this section, we study the effect of splitting classifiers from a more practical perspective. First, we restrict the classifiers over a hypothesis class (e.g., logistic regressions) and analyze the hypothesis class dependent splitting. Second, we consider splitting classifiers in a finite sample regime and study the sample limited splitting.

A. Hypothesis Class Dependent Splitting

We restrict both group-blind and split classifiers over the same hypothesis class and introduce a hypothesis class dependent benefit-of-splitting for quantifying the loss reduction by splitting classifiers.

Definition 5. For a fixed probability distribution $P_{X,Y|S=s}$ with $s\in\{0,1\}$ and a given hypothesis class $\mathcal{H}$, the $\mathcal{H}$-benefit-of-splitting is defined as

$$\begin{aligned}\epsilon_{\text{split}}^{\mathcal{H}} \triangleq &\inf_{h\in\mathcal{H}}\max_{s\in\{0,1\}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right]\\ &-\max_{s\in\{0,1\}}\inf_{h\in\mathcal{H}}\mathbb{E}\left[|h(X)-y_s(X)|\mid S=s\right].\end{aligned} \tag{7}$$

Clearly, the $\mathcal{H}$-benefit-of-splitting maintains the non-maleficence principle $\epsilon_{\text{split}}^{\mathcal{H}}\geq 0$, i.e., given sufficient samples, splitting classifiers will never diminish model accuracy compared to using a group-blind classifier. Next, we provide upper and lower bounds for $\epsilon_{\text{split}}^{\mathcal{H}}$ in order to understand when splitting classifiers brings the most benefit. As before, these bounds rely on three major factors: (i) disagreement between optimal (split) classifiers; (ii) similarity between unlabeled distributions; and (iii) approximation error defined as the smallest loss achieved by split classifiers. In particular, we assume that the last factor is small. This is a common assumption in, e.g., the domain adaptation literature [19] since when the hypothesis class is complex enough, this term will be negligible. Furthermore, one central notion of fairness we follow is non-maleficence (i.e., classifiers should avoid the

causation of harm on any group). When the approximation error is large, neither group-blind classifiers nor splitting classifiers are accurate and "harm" is inevitable. Hence, one should change the hypothesis class first instead of splitting.

Theorem 3. *Let h_s^* be an optimal classifier for group $s \in \{0,1\}$:*

$$h_s^* \in \operatorname*{argmin}_{h\in\mathcal{H}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right].$$

Then we have the following upper and lower bounds for the $\mathcal{H}$-benefit-of-splitting

$$\epsilon_{split}^{\mathcal{H}} \le \min_{s\in\{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right]$$

$$\epsilon_{split}^{\mathcal{H}} \ge \frac{1}{2}\max_{s\in\{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right] - \mathrm{D}_{\mathsf{TV}}(P_0\|P_1) - \frac{3\sum_{s\in\{0,1\}} \mathbb{E}\left[|h_s^*(X) - y_s(X)| \mid S = s\right]}{2}.$$

Proof. See Appendix D-A. □

Analogous to our discussions in Section III, the bounds in Theorem 3 delineate a taxonomy of splitting when both group-blind and split classifiers are restricted over the same hypothesis class: splitting classifiers does not bring much benefit when two groups have similar optimal classifiers; splitting classifiers benefits the most when two groups have similar unlabeled distributions and different optimal classifiers. We further demonstrate this taxonomy of splitting and show how these factors influence the effect of splitting through numerical experiments in Section VI-B.

In contrast to the upper bound for ϵ_{split} (see Theorem 1), the upper bound for $\epsilon_{\text{split}}^{\mathcal{H}}$ does not involve the similarity between the unlabeled distributions. Consequently, when the optimal classifiers are different and the unlabeled distributions are different as well, it is unclear how much benefit splitting classifiers brings. We provide the following example which shows that different hypothesis classes may result in largely different values of the $\mathcal{H}$-benefit-of-splitting. Hence, one must study the effect of splitting on a case-by-case basis for different hypothesis classes.

Example 1. Let two groups' unlabeled distributions and labeling functions be $P_0 \sim \mathcal{N}(-\mu, 1)$, $y_0(x) = \mathbb{I}[x > -\mu]$ and $P_1 \sim \mathcal{N}(\mu, 1)$, $y_1(x) = \mathbb{I}[x < \mu]$, respectively. As μ grows larger, the distance between the unlabeled distributions P_0 and P_1 increases (i.e., $\mathrm{D}_{\mathsf{TV}}(P_0\|P_1) \to 1$ as $\mu \to \infty$). Now we consider the following two hypothesis classes:

- $\mathcal{H}_{\text{threshold}}$ is the class of threshold functions over $\mathbb{R}$: $\mathbb{I}[x > a]$ or $\mathbb{I}[x < b]$.
- $\mathcal{H}_{\text{interval}}$ is the class of intervals over $\mathbb{R}$: $\mathbb{I}[x \in (a, b)]$.

Here, a, b are allowed to be $-\infty$ and $+\infty$, respectively. In both cases, the labeling functions are included in the hypothesis classes and, hence, are optimal classifiers. The disagreement between these optimal classifiers is at least $1/2$:

$$\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right] \ge 1/2, \quad s \in \{0,1\}.$$

The benefit-of-splitting under $\mathcal{H}_{\text{threshold}}$ is $1/2$ as any group-blind classifier incurs at least $1/2$ loss on the disadvantaged group. On the other hand, as μ becomes larger, the benefit-of-splitting under $\mathcal{H}_{\text{interval}}$ is nearly 0 since a group-blind classifier with the form $h^*(x) = \mathbb{I}[x \in (-\mu, \mu)]$ can achieve almost perfect accuracy.

The previous example shows that using a threshold function as a group-blind classifier will always incur an inevitable accuracy trade-off between two groups. On the other hand, if we enrich the hypothesis class to include interval functions, this trade-off can be reconciled. Motivated by this observation, when two groups have different unlabeled distributions and different labeling functions, we conjecture that the $\mathcal{H}$-benefit-of-splitting is determined by the "richness" of the hypothesis class: a more complex hypothesis class can produce a group-blind classifier which mimics the labeling function of each group in the region they lie in, and, hence, this classifier guarantees high accuracy for both groups. We formalize this intuition through the example of feedforward neural networks. Recall that a sigmoidal function [97] (e.g., logistic function) $S : \mathbb{R} \to \mathbb{R}$ is a bounded measurable function which satisfies $S(z) \to 1$ as $z \to +\infty$ and $S(z) \to 0$ as $z \to -\infty$. The hypothesis class associated to feedforward neural networks with one layer of sigmoidal functions has the form

$$\mathcal{H} = \left\{\sum_{i=1}^{k} c_i S(a_i \cdot x + b_i) + c_0 \mid a_i \in \mathbb{R}^d, b_i, c_i \in \mathbb{R}\right\}. \quad (8)$$

In this case, Barron's approximation bounds [97] guarantee that these neural networks can approximate a large class of functions reliably.

Proposition 3. *Consider the hypothesis class $\mathcal{H}$ in* (8)*. If $\mathcal{X} \subset \mathbb{R}^d$ is compact, we have*

$$\epsilon_{split}^{\mathcal{H}} \le \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \times \sqrt{1 - \mathrm{D}_{\mathsf{TV}}(P_0\|P_1)} + \frac{2\mathsf{diam}(\mathcal{X})C}{\sqrt{k}},$$

where $\mathsf{diam}(\mathcal{X}) = \sup_{x,x'\in\mathcal{X}}\|x - x'\|_2$,

$$\begin{aligned} h^*(x) &\triangleq \frac{y_0(x)dP_0(x) + y_1(x)dP_1(x)}{dP_0(x) + dP_1(x)} \\ &= \int_{\mathbb{R}^d} \exp(iwx)\widetilde{h^*}(w)dw \end{aligned} \quad (9)$$

for some complex-valued function $\widetilde{h^*}$, and $C \triangleq \int_{\mathbb{R}^d}\|w\|_2|\widetilde{h^*}(w)|dw$.

Proof. See Appendix D-B. □

Remark 4. The condition in (9) goes back to the seminal work of Barron [97]. By the Fourier inversion theorem, if both h^* and its Fourier transform are integrable, this condition is satisfied. Further situations where (9) holds are discussed in [97, Section IX].

In contrast to Theorem 3, the upper bound for the $\mathcal{H}$-benefit-of-splitting above involves the similarity between the unlabeled distributions (i.e., $\mathrm{D}_{\mathsf{TV}}(P_0\|P_1)$) at the cost of having an additional term which is inversely proportional to the hypothesis class complexity. The intuition behind our proof is that if a data scientist is able to train a neural network

with enough neurons, a group-blind classifier is capable of guaranteeing high accuracy for both groups when their unlabeled distributions are different. Consequently, there is no much room for accuracy improvement by splitting classifiers.

B. Comparison with the Cost-of-Coupling

We compare our notion of the $\mathcal{H}$-benefit-of-splitting with the cost-of-coupling introduced by Dwork *et al.* [13]. We first illustrate the difference between group blind, coupled, split classifiers through the example of logistic regressions:

- a group blind classifier never uses a sensitive attribute as an input: $h(x) = \mathsf{logistic}(w^T x)$;
- a coupled classifier uses a sensitive attribute while sharing other parameters: $h(s, x) = \mathsf{logistic}(w^T x + w_0 s)$;
- split classifiers are a set of classifiers applied to each separate group: $h_s(x) = \mathsf{logistic}(w_s^T x)$.

Now we recast the definition of the cost-of-coupling [13] using our notation.

Definition 6 ([13])**.** Let $\mathcal{H}_{\mathrm{C}}$ be a hypothesis class which contains coupled classifiers from a finite set $\mathcal{S} \times \mathcal{X}$ to $[0, 1]$. For a given loss function $\ell(\cdot, \cdot)$, the cost-of-coupling is defined as

$$\max_{P_{S,X,Y}} \Big\{ \min_{h \in \mathcal{H}_{\mathrm{C}}} L(h) - \min_{\substack{h_s \in \mathcal{H}_{\mathrm{C}} \\ \text{for } s \in \mathcal{S}}} L(\{h_s\}_{s \in \mathcal{S}}) \Big\},$$

where the maximum is over all distributions on $\mathcal{S} \times \mathcal{X} \times \{0, 1\}$ and $L(h) \triangleq \mathbb{E}\left[\ell(Y, h(S, X))\right]$, $L(\{h_s\}_{s \in \mathcal{S}}) \triangleq \mathbb{E}\left[\ell(Y, h_S(S, X))\right]$.

There are two important differences between the $\mathcal{H}$-benefit-of-splitting (see Definition 5) and the cost-of-coupling [13]. First, our notion quantifies the gain in accuracy by using split classifiers rather than a group-blind classifier. In contrast, the cost-of-coupling compares coupled classifiers with split classifiers which both take a sensitive attribute as an input. Second, the cost-of-coupling is a worst-case quantity as it maximizes over all distributions. By allowing our notion to rely on the data distribution, Definition 5 captures more intricate scenarios for characterizing the benefit of splitting classifiers. Furthermore, by taking the maximum over all distributions, we recover an analogous result of Theorem 2 in [13].

Corollary 1. *There exists a probability distribution $Q_{S,X,Y}$ whose $\mathcal{H}$-benefit-of-splitting is at least $1/2$ under*

1) *Linear predictors: $\mathcal{H} = \{\mathbb{I}[w^T x \geq 0] \mid w \in \mathbb{R}^d\}$;*
2) *Decision trees: $\mathcal{H}$ is the set of binary decision trees.*

Furthermore, under this hypothetical probability distribution $Q_{S,X,Y}$, no matter which group-blind classifier $h \in \mathcal{H}$ is used, there is always a group $s \in \{0, 1\}$ such that $\mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right] \geq 1/2$.

The proof technique used for this corollary can be extended to many other models (e.g., kernel methods or neural networks) and we defer its proof to Appendix D-C.

C. Sample Limited Splitting

Consider the following scenario. A data scientist has access to finitely many samples and she/he solves an empirical risk optimization in order to obtain an optimal group-blind classifier or a set of optimal split classifiers. When these classifiers are deployed on new fresh samples, a natural question is whether the optimal split classifiers still outperform the group-blind classifier. We introduce the sample-limited-splitting which quantifies the effect of splitting classifiers within this finite sample regime.

Definition 7. For a given hypothesis class $\mathcal{H}$ and n_s i.i.d. samples $\{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ from group $s \in \{0, 1\}$, let $\hat{h}^*$ and $\{\hat{h}_s^*\}_{s \in \{0,1\}}$ be optimal group-blind and split classifiers for the empirical ℓ_1 loss, respectively:

$$\hat{h}^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \max_{s \in \{0,1\}} \frac{\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|}{n_s}, \tag{10}$$

$$\hat{h}_s^* \in \operatorname*{argmin}_{h \in \mathcal{H}} \frac{\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|}{n_s}, \quad s \in \{0, 1\}. \tag{11}$$

The sample-limited-splitting is defined as

$$\begin{aligned} \hat{\epsilon}_{\text{split}} \triangleq \max_{s \in \{0,1\}} \mathbb{E}\left[|\hat{h}^*(X) - y_s(X)| \mid S = s \right] \\ - \max_{s \in \{0,1\}} \mathbb{E}\left[|\hat{h}_s^*(X) - y_s(X)| \mid S = s \right]. \end{aligned} \tag{12}$$

Unlike the benefit-of-splitting or the $\mathcal{H}$-benefit-of-splitting, the sample-limited-splitting is not necessarily non-negative. In other words, with limited amount of samples available, splitting classifiers may not improve accuracy for both groups. In what follows, we provide data-dependent upper and lower bounds for the sample-limited-splitting in order to understand the effect of splitting classifiers in the finite sample regime.

Theorem 4. *Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0, 1\}$ with VC dimension D. If $\hat{h}_s^*$ is a minimizer of the empirical ℓ_1 loss $\sum_{i=1}^{n_s} |h(x_{s,i}) - y_{s,i}|/n_s$ computed via n_s i.i.d. samples $\{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$, then, with probability at least $1 - \delta$,*

$$\begin{aligned} \hat{\epsilon}_{\textit{split}} &\leq \min_{s \in \{0,1\}} \frac{\sum_{i=1}^{n_s} |\hat{h}_1^*(x_{s,i}) - \hat{h}_0^*(x_{s,i})|}{n_s} + \Omega, \\ \hat{\epsilon}_{\textit{split}} &\geq \frac{1}{2} \max_{s \in \{0,1\}} \frac{\sum_{i=1}^{n_s} |\hat{h}_1^*(x_{s,i}) - \hat{h}_0^*(x_{s,i})|}{n_s} \\ &\quad - \mathrm{D}_{\mathsf{TV}}(\hat{P}_0 \| \hat{P}_1) - 3\lambda - \Omega, \end{aligned}$$

where $\hat{P}_s$ is the empirical unlabeled distribution and

$$\lambda \triangleq \frac{1}{2} \left(\frac{\sum_{i=1}^{n_0} |\hat{h}_0^*(x_{0,i}) - y_{0,i}|}{n_0} + \frac{\sum_{i=1}^{n_1} |\hat{h}_1^*(x_{1,i}) - y_{1,i}|}{n_1} \right),$$

$$\Omega \triangleq 4 \max_{s \in \{0,1\}} \sqrt{\frac{2D \log(6n_s) + 2\log(8/\delta)}{n_s}}.$$

Proof. See Appendix D-D. □

Here, the term λ is the (average) training loss and Ω is the complexity term, which is approximately $\sqrt{D/\min\{n_0, n_1\}}$. As shown, the upper and lower bounds for $\hat{\epsilon}_{\text{split}}$ rely on four factors. The first three factors, which also appear in our bounds of the $\mathcal{H}$-benefit-of-splitting (see Theorem 3), are the

disagreement between the (empirically) optimal classifiers, the similarity of the (empirically) unlabeled distributions, and the (empirically) training error. In addition to these factors, our bounds for $\hat{\epsilon}_{\text{split}}$ also depend on the number of samples from each group, especially minority group with less samples, and model complexity (measured by the VC dimension [98]).

VI. Numerical Experiments

We illustrate the theoretical results presented in this paper through experiments. In Section IV, we presented an algorithm (Algorithm 1) for computing the benefit-of-splitting. In particular, when the data distribution is known, this algorithm provably converges to the exact value of the benefit-of-splitting. To evaluate Algorithm 1, we conduct experiments on a synthetic example where both the data distribution and the values of the benefit-of-splitting are known. In Section V, we characterized a taxonomy of splitting when classifiers are restricted over a hypothesis class. We demonstrate this taxonomy of splitting through experiments on 40 real-world datasets.

A. Synthetic Datasets

We introduced the FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ in Section II-B and proposed an efficient procedure for computing this quantity (Algorithm 1). Here, we validate Algorithm 1 through experiments on synthetic datasets. For a fixed parameter $\theta \in [0, \pi/2]$, let two groups' unlabeled distributions be zero-mean Gaussian distributions with different covariance matrices: $P_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$ and $P_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$ where

$$\mathbf{\Sigma}_0 = \begin{pmatrix} 0.5\cos(\theta)^2 + 1 & 0.5\sin(\theta)\cos(\theta) \\ 0.5\sin(\theta)\cos(\theta) & 0.5\sin(\theta)^2 + 1 \end{pmatrix},$$
$$\mathbf{\Sigma}_1 = \begin{pmatrix} 0.5\cos(\theta)^2 + 1 & -0.5\sin(\theta)\cos(\theta) \\ -0.5\sin(\theta)\cos(\theta) & 0.5\sin(\theta)^2 + 1 \end{pmatrix}.$$

The distributions P_0 and P_1 correspond to θ counterclockwise and clockwise rotation of the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathsf{diag}(1.5, 1))$. Furthermore, let the labeling functions be

$$y_0(x) = \begin{cases} 1 & \text{if } (-\sin(\theta), \cos(\theta)) \cdot x > 0 \\ 0 & \text{otherwise,} \end{cases}$$
$$y_1(x) = \begin{cases} 1 & \text{if } (\sin(\theta), \cos(\theta)) \cdot x > 0 \\ 0 & \text{otherwise.} \end{cases}$$

The left-hand side of Figure 2 displays the level sets of P_0 as well as its labeling function.

In this synthetic example, $\epsilon_{\text{split,FER}}$ has a closed-form expression: $\epsilon_{\text{split,FER}} = 2\Pr(X \in \mathcal{A} | S = 0)$ where the set $\mathcal{A} \triangleq \{x = (x_1, x_2) \in \mathbb{R}^2 \mid y_1(x) = 1, x_2 < 0\}$ (see Appendix E-A for a proof). When $\theta = 0$, two groups share the same unlabeled distribution (i.e., $P_0 = P_1$) and the same labeling function (i.e., $y_0 = y_1$). Hence, there is no benefit of splitting classifiers: $\epsilon_{\text{split,FER}} = 0$. On the other hand, when $\theta = \pi/2$, two groups have the same unlabeled distribution but completely different labeling functions. Splitting classifiers achieves the most benefit: $\epsilon_{\text{split,FER}} = 0.5$.

By varying the values of θ and drawing 10k samples from each group, we compare the true values of $\epsilon_{\text{split,FER}}$ with the outputs from Algorithm 1 as well as other empirical approximations. Recall that Algorithm 1 requires a conditional distribution $\Pr(S = s \mid X = x)$ and the labeling functions y_0 and y_1. Since the conditional distribution and labeling functions are known in this synthetic example, we feed their explicit forms into Algorithm 1 for computing $\epsilon_{\text{split,FER}}$ (orange curve in Figure 2 Right). In practice, the conditional distribution and labeling functions are unknown, so we also train a Naive Bayes classifier [99] to approximate $\Pr(S = s \mid X = x)$ and two linear support-vector machine (SVM) classifiers [99] to approximate the labeling functions. By feeding these binary classifiers into Algorithm 1, another approximation of $\epsilon_{\text{split,FER}}$ is output (red curve in Figure 2 Right). Furthermore, we compute $\epsilon_{\text{split,FER}}$ empirically by training optimal group-blind and split classifiers via logistic regression, linear SVM, or Naive Bayes classifier. Computing the false error rate reduction leads to three empirical approximations of $\epsilon_{\text{split,FER}}$.

As shown in Figure 2, when Algorithm 1 has access to the explicit forms of $\Pr(S = s \mid X = x)$, y_0, and y_1, it accurately recovers $\epsilon_{\text{split,FER}}$. This is remarkable since even with the knowledge of the underlying distributions, it is unclear how to compute $\epsilon_{\text{split,FER}}$ directly from its definition. We also observe that Algorithm 1 applied to binary classifiers outputs more accurate approximation of $\epsilon_{\text{split,FER}}$ than the approximations produced by using logistic regression, linear SVM, or Naive Bayes classifier.

To summarize, we conclude that (i) when the underlying distribution is known, Algorithm 1 can produce the precise values of $\epsilon_{\text{split,FER}}$ and has convergence guarantees; (ii) when Algorithm 1 is fed with binary classifiers, it produces reliable approximation of $\epsilon_{\text{split,FER}}$; (iii) computing the FER-benefit-of-splitting empirically by training optimal classifiers could incur high approximation errors.

B. Real-world Datasets

In Section V, we analyzed the effect of splitting classifiers when both group-blind and split classifiers are restricted over the same hypothesis class. The bounds in Theorem 3 reveal two main factors that could determine this effect: disagreement between optimal classifiers and similarity between unlabeled distributions. Here we demonstrate how these two factors influence the effect of splitting through experiments on 40 real-world datasets, collected from OpenML [22].

a) Setup: We preprocess all 40 datasets by adopting the procedure described in [13]. All categorical features are transformed into binary by assigning the most frequent object to 1 and the rest of the objects to 0. The first binary feature is selected as the sensitive attribute and, hence, these datasets are "semi-synthetic". We truncate the datasets so that each group contains at most 10k data points. In each dataset, there are at least 8k data points per group, minimizing the effect of potential lack of samples per group.

b) Implementation: We obtain optimal split classifiers via training a logistic regression model with the `LIBLINEAR` solver [100], fitting the model by drawing samples from each group. Since an optimal group-blind classifier is a minimizer of $\min_{h \in \mathcal{H}} \max_{w \in [0,1]} wL_0(h) + (1-w)L_1(h)$ where $L_s(h)$ is the loss of a classifier h on group $s \in \{0, 1\}$, we solve this

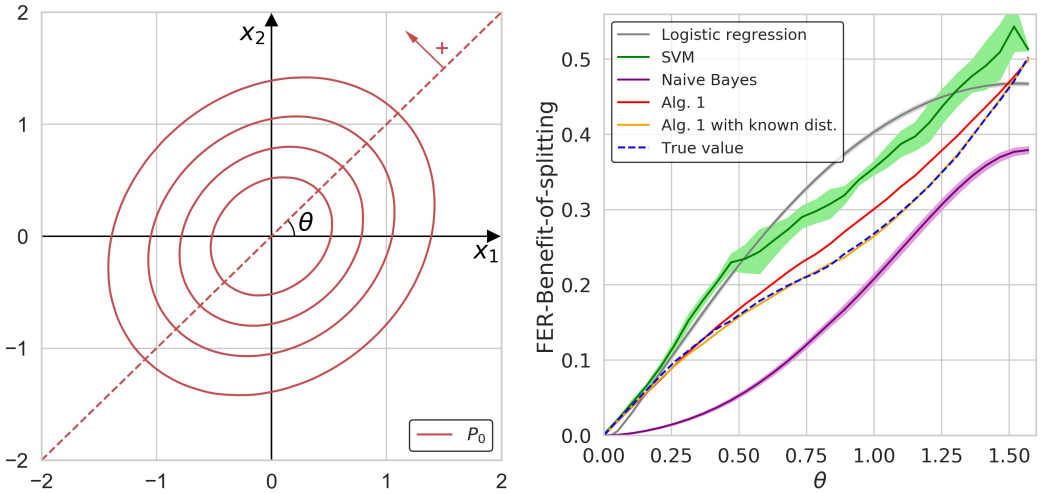

Fig. 2: We demonstrate the performance of Algorithm 1 for computing the FER-benefit-of-splitting $\epsilon_{\text{split,FER}}$ on synthetic datasets. Left: the ellipses are the level sets of the unlabeled distribution P_0 and the dash line is the labeling function y_0 with a arrow indicating the region where points are labeled as $+$. Right: $\epsilon_{\text{split,FER}}$ computed by different approaches along with its true values. As shown, when the underlying data distribution is known, the approximation of $\epsilon_{\text{split,FER}}$ produced by Algorithm 1 (orange curve) recovers its true values (blue dash curve). When the underlying distribution is unknown, we train binary classifiers and feed them into Algorithm 1. The approximation of $\epsilon_{\text{split,FER}}$ produced by Algorithm 1 is depicted as red curve. Finally, we compute $\epsilon_{\text{split,FER}}$ empirically by training optimal group-blind and split classifiers via (i) logistic regression (gray curve), (ii) linear SVM (green curve), (iii) Naive Bayes classifier (purple curve). We use 5-fold cross validation for training these optimal classifiers and plot the standard deviation as shaded region. As shown, the approximations of $\epsilon_{\text{split,FER}}$ produced by Algorithm 1 outperform all three empirical approximations of $\epsilon_{\text{split,FER}}$.

optimization approximately by considering its dual formula $\max_{w\in[0,1]} \min_{h\in\mathcal{H}} wL_0(h) + (1-w)L_1(h)$ and use 5-fold cross validation to tune the parameter w therein. Although this procedure of training group-blind classifier needs access to data points' sensitive attribute, it does not violate group-blindness [85] because the output classifier does not use the sensitive attribute as an input when deploying on new data. In addition to logistic regressions, we repeat this experiment by training decision tree classifiers with depth 7. The disagreement between optimal classifiers is calculated by applying the optimal split classifiers on each data point and computing the discrepancy. We estimate the total variation distance between unlabeled distributions by applying the procedures introduced in [101] (see Appendix E-B for more details).

c) Result: In Figure 3, we illustrate the taxonomy of splitting delineated by our bounds in Theorem 3. We restrict the hypothesis class to be logistic regression (Figure 3 Left) or to be decision trees with depth 7 (Figure 3 Right). Each dot in the figures represents a dataset with its corresponding ID number in the OpenML dataset. The color captures the loss reduction by using the optimal split classifiers compared to deploying the optimal group-blind classifier (red means splitting has more benefit and blue means splitting does not bring much benefit). The location of each dot is determined by the two factors: disagreement between optimal classifiers (y-axis) and total variation distance between unlabeled distributions (x-axis).

- The upper bound in Theorem 3 indicates that splitting does not bring much benefit when the optimal classifiers are similar. As shown in Figure 3, all datasets which are below the horizontal dash line have small benefit by splitting classifiers (i.e., dots are blue).
- The lower bound in Theorem 3 indicates that splitting benefits model performance when the optimal classifiers are different and the unlabeled distributions are similar. As shown in Figure 3, there are two datasets (ID 122 and 1169) which are in the yellow region and they all achieve large benefit from splitting classifiers.
- When both the optimal classifiers and the unlabeled distributions are different, the effect of splitting classifiers can not be determined by the bounds in Theorem 3. As shown in Figure 3, the datasets in the grey region could have either large benefit by splitting classifiers or limited benefit. Furthermore, we have conjectured (see Section V-A) that in this case a more complex hypothesis class leads to less benefit from splitting classifiers. This is further evidenced in the experiments: when both group-blind and split classifiers are logistic regressions (Figure 3 Left), the datasets which are in the grey region all achieve non-trivial benefit by splitting classifiers. In contrast, when decision trees are used (Figure 3 Right), there are datasets (e.g., ID 1240) in the grey region which achieve a limited amount of benefit by splitting.

VII. Conclusion and Future Work

Split classifiers should only be considered when it is fair, ethical, and legal to do so, and when it does not result in harm to any underlying group. Eliminating disparate treatment does not necessarily lead to a group-fair classifier. On the one hand, a sensitive attribute could correlate with other proxy variables which are used for decision making [14], [56]. On the other hand, the sensitive attribute can be an important feature for the prediction task [86], [89]. In the latter case, using a group-

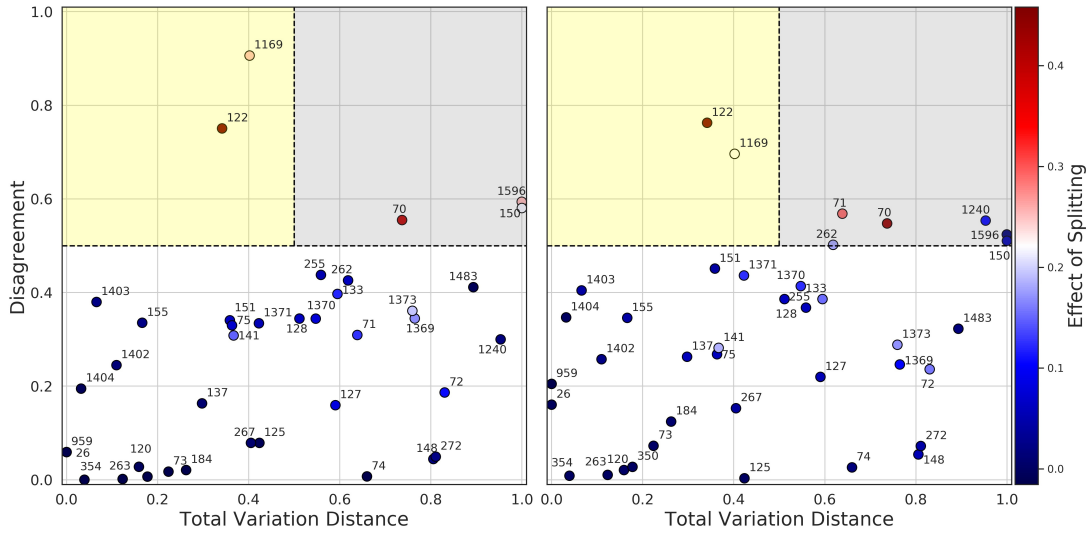

Fig. 3: We demonstrate how the effect of splitting classifiers is determined by the two factors: disagreement between optimal classifiers (y-axis) and total variation distance between unlabeled distributions (x-axis). We restrict both group-blind and split classifiers to logistic regression classifiers (left) or decision tree classifiers (right). Each dot represents a dataset in OpenML [22] with color indicating the effect of splitting classifiers compared to using a group-blind classifier and texts indicating dataset ID. Our upper and lower bounds in Theorem 3 reveal a taxonomy of splitting where splitting does not bring much benefit (white region); splitting brings the most benefit (yellow region); or splitting has undetermined effect (grey region).

blind classifier for achieving treatment parity may lead to an unfavorable accuracy trade-off.

Motivated by the above discussion, we investigated the following fundamental question: when disparate treatment is allowed, is it beneficial to incorporate the sensitive attribute as an input feature in order to improve a classifier's performance? Due to the bias-variance trade-off, in practice, the answer will depend on the number of samples available for training the model and the complexity of the hypothesis class. In this paper, we focused on an information-theoretic regime where the underlying data distribution is known—or infinitely many data points are available—and the hypothesis class is unrestricted. To evaluate the potential gain in average performance from allowing a classifier to exhibit disparate treatment, we compare split classifiers with group-blind classifiers and characterize precise conditions where splitting classifiers achieves the most benefit. Our results show that—in this narrow information-theoretic regime—splitting classifiers follows the non-maleficence principle and allows a data scientist to deploy more accurate and suitable models for each group. However, the use of a sensitive attribute relies on several factors and may even be illegal and unethical for certain tasks [2]. The analysis presented here aims at providing an objective analysis for understanding the benefit (or risk) of splitting only from a theoretical vantage point.

There are two open questions that deserve further exploration. First, our bounds indicate that the difference in underlying data distributions between groups, the number of samples, and the hypothesis class can all influence the effect of splitting classifiers. Nonetheless, we believe that there are more factors that play an important role in determining such an effect. For example, a group-blind classifier may perform worse on minority groups due to unbalanced samples in the training process and using split classifiers could potentially reconcile this issue. In a similar vein, the lack of sample diversity (i.e., training datasets do not contain enough samples from minority groups) could affect the performance and generalization of ML models for minority groups. Hence, it is crucial to characterize the impact of sample size and diversity on detecting and reducing discrimination. Second, we introduce the sample-limited-splitting $\hat{\epsilon}_{\text{split}}$ for quantifying the effect of splitting classifiers in the finite sample regime and provide its upper and lower bounds. It would be interesting to characterize precise conditions under which $\hat{\epsilon}_{\text{split}} \geq 0$ (or $\hat{\epsilon}_{\text{split}} \leq 0$).

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Berk Ustun for his valuable input. The authors would also like to thank the anonymous reviewers and the associate editor for their constructive feedback.

APPENDIX A
EXAMPLES OF f-DIVERGENCE

We recall some examples of f-divergence [45] here.

- KL-divergence [102]: $f(x) = x\log(x)$,

$$\mathsf{D}_{\mathsf{KL}}(P\|Q) = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)\mathrm{d}P.$$

- Total variation distance: $f(x) = |x-1|/2$,

$$\mathsf{D}_{\mathsf{TV}}(P\|Q) = \frac{1}{2}\int\left|\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right|\mathrm{d}Q.$$

- Chi-square divergence [103]: $f(x) = (x-1)^2$ or $f(x) = x^2 - 1$,

$$\begin{aligned}\mathsf{D}_{\chi^2}(P\|Q) &= \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2\mathrm{d}Q \\ &= \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2\mathrm{d}Q - 1.\end{aligned}$$

- Jensen-Shannon divergence [104]: $f(x) = x\log(x)/2 - (1+x)\log((1+x)/2)/2$,

$$\mathsf{JS}(P\|Q) = \frac{1}{2}\mathsf{D}_{\mathsf{KL}}\left(P\|\frac{P+Q}{2}\right) + \frac{1}{2}\mathsf{D}_{\mathsf{KL}}\left(Q\|\frac{P+Q}{2}\right).$$

  Note that the Jensen-Shannon divergence is defined in a general form in [104] for $\omega \in [0,1]$

$$\begin{aligned}\mathsf{JS}_\omega(P\|Q) =& \omega\mathsf{D}_{\mathsf{KL}}\left(P\|\omega P + (1-\omega)Q\right) \\ &+ (1-\omega)\mathsf{D}_{\mathsf{KL}}\left(Q\|\omega P + (1-\omega)Q\right).\end{aligned}$$

- E_γ-divergence (also called hockey-stick divergence) [91], [92], [105], [106]: $f(x) = (x-\gamma)_+$ for $\gamma \geq 1$ where $(a)_+ \triangleq \max\{a, 0\}$,

$$E_\gamma(P\|Q) = \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - \gamma\right)_+\mathrm{d}Q.$$

- DeGroot statistical information [93] of order p: $f(x) = \min\{p, 1-p\} - \min\{p, 1-px\}$ for $p \in (0,1)$,

$$\mathcal{I}_p(P\|Q) = \min\{p,1-p\} - \int\min\left\{p, 1-p\frac{\mathrm{d}P}{\mathrm{d}Q}\right\}\mathrm{d}Q.$$

- Marton's divergence [107]: $f(x) = (x-1)^2\mathbb{I}[x<1]$,

$$\begin{aligned}d_2(P\|Q)^2 &= \inf\mathbb{E}\left[\Pr(X\neq Y \mid Y)^2\right] \\ &= \int\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2\mathbb{I}[\frac{\mathrm{d}P}{\mathrm{d}Q} < 1]\mathrm{d}Q,\end{aligned}$$

  where the infimum is taken over all couplings, i.e., joint distributions $P_{X,Y}$ which have marginals $P_X = P$ and $P_Y = Q$, respectively.

We refer the readers to [91], [108] for more examples of f-divergence and their properties.

APPENDIX B
PROOFS FOR SECTION III

A. Proof of Lemma 2

Proof. We first introduce a (measurable) loss function $\ell : [0,1]\times[0,1] \to \mathbb{R}^+ \cup \{\infty\}$ and assume that this loss function satisfies: (i) $\ell(a,a) = 0$ for any $a \in [0,1]$ and (ii) for any $a \in [0,1]$, $\ell(a,\cdot)$ is convex and continuous. The benefit-of-splitting in Definition 2 can be written as

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\mathbb{E}\left[\ell(y_s(X),h(X)) \mid S=s\right] \\ &- \max_{s\in\{0,1\}}\inf_{h:\mathcal{X}\to[0,1]}\mathbb{E}\left[\ell(y_s(X),h(X)) \mid S=s\right].\end{aligned} \tag{13}$$

By taking the ℓ_1 loss $\ell(a,b) = |a-b|$, ℓ_2 loss $\ell(a,b) = (a-b)^2$, and KL loss $\ell(a,b) = \mathsf{D}_{\mathsf{KL}}(a\|b) \triangleq a\log(a/b) + (1-a)\log((1-a)/(1-b))$, respectively, the above quantity becomes $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split},\mathsf{KL}}$. These loss functions all satisfy our above two assumptions. In particular, by our first assumption, one can choose $h(x) = y_s(x)$ which leads to

$$\max_{s\in\{0,1\}}\inf_{h:\mathcal{X}\to[0,1]}\mathbb{E}\left[\ell(y_s(X),h(X)) \mid S=s\right] = 0. \tag{14}$$

Hence, the problem remains providing equivalent expression for the inf-max term

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\mathbb{E}\left[\ell(y_s(X),h(X)) \mid S=s\right] \\ &= \inf_{h:\mathcal{X}\to[0,1]}\sup_{\omega\in[0,1]}\omega\cdot\mathbb{E}\left[\ell(y_0(X),h(X)) \mid S=0\right] \\ &\qquad + (1-\omega)\cdot\mathbb{E}\left[\ell(y_1(X),h(X)) \mid S=1\right].\end{aligned} \tag{15}$$

Next, we use Ky Fan's min-max theorem [24] (see Lemma 1) to swap the positions of infimum and supremum in (15). We start with verifying the assumptions in Ky Fan's min-max theorem. We denote the set of all measurable functions from $\mathcal{X}$ to $[0,1]$ by $\mathcal{L}(\mathcal{X}\to[0,1])$ and introduce a function $F : [0,1]\times\mathcal{L}(\mathcal{X}\to[0,1]) \to \mathbb{R}$

$$\begin{aligned}F(\omega,h) \triangleq\ & \omega\cdot\mathbb{E}\left[\ell(y_0(X),h(X)) \mid S=0\right] \\ &+ (1-\omega)\cdot\mathbb{E}\left[\ell(y_1(X),h(X)) \mid S=1\right].\end{aligned}$$

For every fixed $h \in \mathcal{L}(\mathcal{X}\to[0,1])$, $F(\cdot,h)$ is a linear function. Consequently, $F(\cdot,h)$ is upper semicontinuous and F is concave-like on $[0,1]$. Furthermore, for any $h_1, h_1 \in \mathcal{L}(\mathcal{X}\to[0,1])$, $\lambda\in[0,1]$, and $\omega\in[0,1]$, we have $\lambda h_1 + (1-\lambda)h_2 \in \mathcal{L}(\mathcal{X}\to[0,1])$ and

$$F(\omega,\lambda h_1 + (1-\lambda)h_2) \leq \lambda F(\omega,h_1) + (1-\lambda)F(\omega,h_2)$$

by the convexity of $\ell(a,\cdot)$ for any $a\in[0,1]$. Hence, F is convex-like on $\mathcal{L}(\mathcal{X}\to[0,1])$. Therefore, by Ky Fan's min-max theorem, (15) is equal to

$$\begin{aligned}&\sup_{\omega\in[0,1]}\inf_{h:\mathcal{X}\to[0,1]}\omega\cdot\mathbb{E}\left[\ell(y_0(X),h(X)) \mid S=0\right] \\ &\qquad + (1-\omega)\cdot\mathbb{E}\left[\ell(y_1(X),h(X)) \mid S=1\right].\end{aligned} \tag{16}$$

Now we take any probability distribution P over $\mathcal{X}$ such that P_0 and P_1 are absolutely continuous with respect to P. For

example, one can simply choose $\mathrm{d}P = (\mathrm{d}P_0 + \mathrm{d}P_1)/2$. Then (16) can be written as

$$\sup_{\omega\in[0,1]} \inf_{h:\mathcal{X}\to[0,1]} \int \Big(\omega \cdot \ell(y_0(x), h(x)) \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega)\cdot \ell(y_1(x), h(x)) \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\Big)\mathrm{d}P(x). \tag{17}$$

Next, we prove that the infimum and the integer in (17) can be interchanged. For a fixed $\omega \in [0,1]$, we introduce a function $f : \mathcal{X} \times [0,1] \to \mathbb{R}$

$$f(x,\bar{h}) \triangleq \omega \cdot \ell(y_0(x),\bar{h}) \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega)\cdot \ell(y_1(x),\bar{h}) \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}$$

and aim at proving

$$\inf_{h:\mathcal{X}\to[0,1]} \int f(x,h(x))\mathrm{d}P(x) = \int \inf_{\bar{h}\in[0,1]} f(x,\bar{h})\mathrm{d}P(x). \tag{18}$$

Since $f(\cdot,\bar{h})$ is measurable and $f(x,\cdot)$ is continuous, f is a Carathéodory function (see Section 4.10 in [109]). Hence, by the measurable maximum theorem (see Theorem 18.19 in [109]), the mapping

$$x \to \inf_{\bar{h}\in[0,1]} f(x,\bar{h})$$

is measurable and the argmin correspondence (i.e., set-valued function)

$$\mathcal{H}^*(x) \triangleq \left\{\bar{h}^* \in [0,1] \mid f(x,\bar{h}^*) = \inf_{\bar{h}\in[0,1]} f(x,\bar{h})\right\}$$

is also measurable and admits a measurable selector. We denote this selector by $h^* : \mathcal{X} \to [0,1]$ and, by definition, it satisfies $h^*(x) \in \mathcal{H}^*(x)$ for all $x \in \mathcal{X}$. Now we are ready to prove (18). One direction LHS $\geq$ RHS can be obtained directly since for any $h : \mathcal{X} \to [0,1]$

$$\int f(x,h(x))\mathrm{d}P(x) \geq \int \inf_{\bar{h}\in[0,1]} f(x,\bar{h})\mathrm{d}P(x).$$

By the definition of $h^*(x)$,

$$\begin{aligned}\text{RHS} &= \int f(x,h^*(x))\mathrm{d}P(x)\\ &\geq \inf_{h:\mathcal{X}\to[0,1]} \int f(x,h(x))\mathrm{d}P(x) = \text{LHS}.\end{aligned}$$

Therefore, the equality in (18) holds and (17) becomes

$$\sup_{\omega\in[0,1]} \int \Big(\omega \cdot \ell(y_0(x), h^*(x)) \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} + (1-\omega)\cdot \ell(y_1(x), h^*(x)) \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\Big)\mathrm{d}P(x). \tag{19}$$

Hence, our last step is to compute the function h^* for the loss functions of interest. If the loss function is ℓ_1, then $\operatorname{argmin}_{\bar{h}\in[0,1]} f(x,\bar{h})$ is equal to

$$\operatorname*{argmin}_{\bar{h}\in[0,1]} \left\{\omega \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)}|\bar{h} - y_0(x)| + (1-\omega)\frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}|\bar{h} - y_1(x)|\right\}.$$

For a fixed $\omega \in [0,1]$, the optimal classifier is

$$h^*(x) = \begin{cases} y_0(x) & \text{if } \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq \frac{1-\omega}{\omega} \\ y_1(x) & \text{otherwise.}\end{cases}$$

By substituting the optimal classifier and ℓ_1 loss into (19), we get the desired equivalent expression of $\epsilon_{\text{split},1}$:

$$\sup_{\omega\in[0,1]} (1-\omega)\int_{\mathcal{A}_\omega} |y_1(x) - y_0(x)|\mathrm{d}P_1(x) + \omega \int_{\mathcal{A}_\omega^c} |y_1(x) - y_0(x)|\mathrm{d}P_0(x),$$

where $\mathcal{A}_\omega \triangleq \left\{x \mid \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq \frac{1-\omega}{\omega}\right\}$. Similarly, when ℓ_2 loss is used, the optimal classifier becomes

$$h^*(x) = \frac{\omega y_0(x)\mathrm{d}P_0(x) + (1-\omega)y_1(x)\mathrm{d}P_1(x)}{\omega \mathrm{d}P_0(x) + (1-\omega)\mathrm{d}P_1(x)}, \tag{20}$$

which leads to the equivalent expression of $\epsilon_{\text{split},2}$:

$$\sup_{\omega\in[0,1]} \omega(1-\omega)\int \frac{(y_1(x) - y_0(x))^2 \mathrm{d}P_0(x)\mathrm{d}P_1(x)}{\omega \mathrm{d}P_0(x) + (1-\omega)\mathrm{d}P_1(x)}.$$

When the KL-loss is used, the optimal classifier h^* has expression in (20) as well. Consequently, we have the equivalent expression of $\epsilon_{\text{split},\mathsf{KL}}$:

$$\sup_{\omega\in[0,1]} \omega\mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_0(X)\|h^*(X)) \mid S=0\right] + (1-\omega)\mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_1(X)\|h^*(X)) \mid S=1\right]. \tag{21}$$

This expression can be further simplified by using the chain rule of KL-divergence:

$$\begin{aligned}&\mathrm{D}_{\mathsf{KL}}(Q_{X,Y}\|R_{X,Y})\\ &= \mathrm{D}_{\mathsf{KL}}(Q_{Y|X}\|R_{Y|X} \mid Q_X) + \mathrm{D}_{\mathsf{KL}}(Q_X\|R_X).\end{aligned}$$

By taking $\mathrm{d}Q_X = \mathrm{d}P_s$, $\mathrm{d}R_X = w\mathrm{d}P_0 + (1-w)\mathrm{d}P_1$, $Q_{Y|X}(1|x) = y_s(x)$, and $R_{Y|X}(1|x) = h^*(x)$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_s(X)\|h^*(X)) \mid S=s\right]\\ &= \mathrm{D}_{\mathsf{KL}}\left(P_{X,Y|S=s}\|\omega P_{X,Y|S=0} + (1-\omega)P_{X,Y|S=1}\right)\\ &\quad - \mathrm{D}_{\mathsf{KL}}\left(P_s\|\omega P_0 + (1-\omega)P_1\right).\end{aligned}$$

Substituting this into (21) gives

$$\epsilon_{\text{split},\mathsf{KL}} = \sup_{\omega\in[0,1]} \mathsf{JS}_\omega(P_{X,Y|S=0}\|P_{X,Y|S=1}) - \mathsf{JS}_\omega(P_0\|P_1),$$

where $\mathsf{JS}_\omega(\cdot\|\cdot)$ is the Jensen-Shannon divergence (see Appendix A for its definition). □

B. Proof of Theorem 1

We divide the proof of Theorem 1 into three independent steps. First, we prove the upper bounds for $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split},\mathsf{KL}}$ in a unified way. Then we prove the lower bounds for $\epsilon_{\text{split},1}$ and $\epsilon_{\text{split},\mathsf{KL}}$ using Lemma 2. Finally, we prove the lower bound for $\epsilon_{\text{split},2}$ by leveraging the proof techniques of Brown-Low's two-points lower bound [23].

Proof. Note that (14) implies in the information-theoretic regime, optimal split classifiers can always achieve perfect performance. Specifically, one can select labeling functions y_0

and y_1 as split classifiers which have zero loss on each group. Hence, the problem remains upper bounding the performance of the optimal group-blind classifier. To achieve this goal, we consider two special group-blind classifiers:

$$h^*(x) = \frac{\mathrm{d}P_0(x)}{2\mathrm{d}P(x)} y_0(x) + \frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)} y_1(x), \tag{22}$$

$$h^{**}(x) = \frac{1}{2}(y_0(x) + y_1(x)), \tag{23}$$

where $\mathrm{d}P = (\mathrm{d}P_0 + \mathrm{d}P_1)/2$. In what follows, we upper bound the performance of the group-blind classifiers in (22) and (23) and these bounds will be naturally translated into the upper bounds of $\epsilon_{\text{split},1}$, $\epsilon_{\text{split},2}$, and $\epsilon_{\text{split},\mathsf{KL}}$, respectively.

We upper bound $\epsilon_{\text{split},1}$ by using the group-blind classifier h^* in (22).

$$\begin{aligned}
\epsilon_{\text{split},1} &= \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right] \\
&\le \max_{s\in\{0,1\}} \mathbb{E}\left[|h^*(X) - y_s(X)| \mid S = s\right] \\
&= \int |y_1(x) - y_0(x)| \frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)} \mathrm{d}P_0(x).
\end{aligned} \tag{24}$$

By the Cauchy-Schwarz inequality, we can further upper bound (24) by

$$\sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right] \cdot \int \left(\frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x)}. \tag{25}$$

Furthermore, we have

$$\begin{aligned}
&\int \left(\frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x) \\
&= \frac{1}{4} \int \left(\frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right)^2 \frac{\mathrm{d}P_0(x)}{\mathrm{d}P(x)} \mathrm{d}P(x) \\
&= \frac{1}{4} \int \left(\frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right)^2 \left(2 - \frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)}\right) \mathrm{d}P(x).
\end{aligned} \tag{26}$$

Since $\frac{1}{4}x^2(2-x) \le 1 - |x-1|$ holds for any $x \ge 0$, the RHS of (26) can be upper bounded by

$$\begin{aligned}
1 - \int \left|\frac{\mathrm{d}P_1(x)}{\mathrm{d}P(x)} - 1\right| \mathrm{d}P(x) &= 1 - \frac{1}{2}\int |\mathrm{d}P_1(x) - \mathrm{d}P_0(x)| \\
&= 1 - \mathrm{D}_{\mathsf{TV}}(P_0 \| P_1).
\end{aligned} \tag{27}$$

Combining (24–27) gives an upper bound of $\epsilon_{\text{split},1}$:

$$\sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right]} \cdot \sqrt{1 - \mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)}.$$

By symmetry, we can further tighten this upper bound by replacing it with

$$\min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \cdot \sqrt{1 - \mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)}.$$

On the other hand, using the classifier h^{**} in (23) leads to an alternative upper bound

$$\epsilon_{\text{split},1} \le \frac{1}{2} \max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right].$$

Similarly, we can upper bound $\epsilon_{\text{split},2}$ by using the classifier h^* in (22)

$$\begin{aligned}
\epsilon_{\text{split},2} &\le \max_{s\in\{0,1\}} \mathbb{E}\left[(h^*(X) - y_s(X))^2 \mid S = s\right] \\
&\le \int (y_1(x) - y_0(x))^2 \left(\frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_0(x) \\
&\quad + \int (y_1(x) - y_0(x))^2 \left(\frac{\mathrm{d}P_0(x)}{2\mathrm{d}P(x)}\right)^2 \mathrm{d}P_1(x) \\
&= \int (y_1(x) - y_0(x))^2 \frac{\mathrm{d}P_1(x)}{2\mathrm{d}P(x)} \mathrm{d}P_0(x),
\end{aligned} \tag{28}$$

where the second inequality uses the fact that $\max\{a, b\} \le a + b$. By the Cauchy-Schwarz inequality and (26), (27), we can further upper bound (28) by

$$\sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^4 \mid S = 0\right]} \cdot \sqrt{1 - \mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)}.$$

By symmetry, we can further tighten this upper bound by replacing it with

$$\min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^4 \mid S = s\right]} \cdot \sqrt{1 - \mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)}.$$

On the other hand, using the classifier h^{**} in (23) leads to an alternative upper bound for $\epsilon_{\text{split},2}$.

We repeat the same strategy and upper bound $\epsilon_{\text{split},\mathsf{KL}}$ by using the classifier h^* in (22)

$$\begin{aligned}
\epsilon_{\text{split},\mathsf{KL}} &\le \max_{s\in\{0,1\}} \mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_s(X) \| h^*(X)) \mid S = s\right] \\
&\le \mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_0(X) \| h^*(X)) \mid S = 0\right] \\
&\quad + \mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_1(X) \| h^*(X)) \mid S = 1\right].
\end{aligned}$$

Recall the chain rule of KL-divergence

$$\begin{aligned}
&\mathrm{D}_{\mathsf{KL}}(Q_{X,Y} \| R_{X,Y}) \\
&= \mathrm{D}_{\mathsf{KL}}(Q_{Y|X} \| R_{Y|X} \mid Q_X) + \mathrm{D}_{\mathsf{KL}}(Q_X \| R_X).
\end{aligned}$$

By taking $\mathrm{d}Q_X = \mathrm{d}P_s$, $\mathrm{d}R_X = \mathrm{d}P$, $Q_{Y|X}(1|x) = y_s(x)$, and $R_{Y|X}(1|x) = h^*(x)$ and noticing the definition of h^* in (22), we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{D}_{\mathsf{KL}}(y_s(X) \| h^*(X)) \mid S = s\right] \\
&= \mathrm{D}_{\mathsf{KL}}\left(P_{X,Y|S=s} \| \frac{P_{X,Y|S=0} + P_{X,Y|S=1}}{2}\right) \\
&\quad - \mathrm{D}_{\mathsf{KL}}\left(P_s \| \frac{P_0 + P_1}{2}\right).
\end{aligned} \tag{29}$$

Hence,

$$\epsilon_{\text{split},\mathsf{KL}} \le 2\mathsf{JS}(P_{X,Y|S=0} \| P_{X,Y|S=1}) - 2\mathsf{JS}(P_0 \| P_1),$$

where $\mathsf{JS}(\cdot\|\cdot)$ is the Jensen-Shannon divergence. On the other hand, taking the classifier h^{**} in (23) gives an alternative upper bound for $\epsilon_{\text{split},\mathsf{KL}}$. □

We proceed to prove the lower bounds of $\epsilon_{\text{split},1}$ and $\epsilon_{\text{split},\mathsf{KL}}$.

Proof. Recall that $\mathcal{A}_{0.5} \triangleq \left\{x \in \mathcal{X} \mid \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} \geq 1\right\}$. By Lemma 2, we can lower bound $\epsilon_{\text{split},1}$ by

$$\begin{aligned}
&\frac{1}{2}\Big(\int_{\mathcal{A}_{0.5}} |y_1(x) - y_0(x)| \mathrm{d}P_1(x) \\
&\quad + \int_{\mathcal{A}_{0.5}^c} |y_1(x) - y_0(x)| \mathrm{d}P_0(x)\Big) \\
&= \frac{1}{2}\Big(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] \\
&\qquad - \int_{\mathcal{A}_{0.5}^c} |y_1(x) - y_0(x)| (\mathrm{d}P_1(x) - \mathrm{d}P_0(x))\Big) \\
&= \frac{1}{2}\Big(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] \\
&\qquad - \int |y_1(x) - y_0(x)| \left(1 - \frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)}\right)_+ \mathrm{d}P_1(x)\Big) \\
&\geq \frac{1}{2}\Big(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] \\
&\qquad - \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 1\right]} \cdot d_2(P_0 \| P_1)\Big),
\end{aligned}$$

where $d_2(P_0\|P_1)$ is Marton's divergence. By symmetry, one can obtain the desired lower bound of $\epsilon_{\text{split},1}$. Finally, the lower bound of $\epsilon_{\text{split,KL}}$ follows directly from Lemma 2. □

Before getting to the lower bound of $\epsilon_{\text{split},2}$, we prove a useful lemma.

Lemma 3. *For any measurable classifier $h : \mathcal{X} \to [0,1]$ and constant $0 \leq \epsilon < \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right]$, if the condition $\mathbb{E}\left[(h(X) - y_0(X))^2 \mid S = 0\right] \leq \epsilon$ holds, then $\mathbb{E}\left[(h(X) - y_1(X))^2 \mid S = 1\right] \geq (A - B\sqrt{\epsilon})^2$, where*

$$A \triangleq \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right], \tag{30a}$$

$$B \triangleq \sqrt{\mathrm{D}_{\chi^2}(P_1\|P_0) + 1}. \tag{30b}$$

Proof. Consider a convex optimization problem

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

Computing the Gateaux derivative of the Lagrange multiplier gives the following optimal conditions (see Theorem 6.6.1 in [110]),

$$(h(x) - y_1(x))\mathrm{d}P_1(x) + \lambda(h(x) - y_0(x))\mathrm{d}P_0(x) = 0, \tag{31}$$

$$\lambda\left(\int (h(x) - y_0(x))^2 \mathrm{d}P_0(x) - \epsilon\right) = 0, \tag{32}$$

$$\lambda \geq 0, \tag{33}$$

which provides the optimal classifier

$$h^*(x) = \frac{y_1(x)\mathrm{d}P_1(x) + \lambda y_0(x)\mathrm{d}P_0(x)}{\mathrm{d}P_1(x) + \lambda \mathrm{d}P_0(x)}.$$

We denote $r(x) \triangleq \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}$ and simplify the expression of the optimal classifier

$$h^*(x) = \frac{y_1(x)r(x) + \lambda y_0(x)}{r(x) + \lambda}. \tag{34}$$

If $\lambda = 0$, then $h^*(x) = y_1(x)$ and, consequently,

$$\begin{aligned}
&\mathbb{E}\left[(h^*(X) - y_0(X))^2 \mid S = 0\right] \\
&= \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right].
\end{aligned}$$

However, this contradicts our assumptions:

$$\begin{aligned}
&\mathbb{E}\left[(h^*(X) - y_0(X))^2 \mid S = 0\right] \leq \epsilon, \\
&\epsilon < \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right].
\end{aligned}$$

Hence, we have $\lambda > 0$. In this case, (32) and (34) imply

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

We simplify the expression and obtain

$$\int r(x)^2 \left(\frac{y_1(x) - y_0(x)}{r(x) + \lambda}\right)^2 \mathrm{d}P_0(x) = \epsilon. \tag{35}$$

Now we consider lower bounding the quantity $\mathbb{E}\left[(h^*(X) - y_1(X))^2 \mid S = 1\right]$. By its definition and the expression of the optimal classifier (34), we have

$$\begin{aligned}
&\mathbb{E}\left[(h^*(X) - y_1(X))^2 \mid S = 1\right] \\
&= \int \left(\frac{y_1(x)r(x) + \lambda y_0(x)}{r(x) + \lambda} - y_1(x)\right)^2 \mathrm{d}P_1(x) \\
&= \int \left(\frac{\lambda(y_1(x) - y_0(x))}{r(x) + \lambda}\right)^2 \mathrm{d}P_1(x) \\
&\geq \left(\int \frac{\lambda|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x)\right)^2 \\
&= \Big(\int |y_1(x) - y_0(x)| \mathrm{d}P_1(x) \\
&\qquad - \int \frac{r(x)|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x)\Big)^2,
\end{aligned} \tag{36}$$

where the only inequality is due to the Cauchy-Schwarz inequality. Furthermore, by the Cauchy-Schwarz inequality again and (35), we have

$$\begin{aligned}
&\int \frac{r(x)|y_1(x) - y_0(x)|}{r(x) + \lambda} \mathrm{d}P_1(x) \\
&\leq \sqrt{\int r(x)^2 \left(\frac{y_1(x) - y_0(x)}{r(x) + \lambda}\right)^2 \mathrm{d}P_0(x) \int r(x) \mathrm{d}P_1(x)} \\
&= \sqrt{\epsilon \mathbb{E}\left[r(X) \mid S = 1\right]}.
\end{aligned} \tag{37}$$

Recall that $r(x) = \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}$. Hence,

$$\begin{aligned}
\mathbb{E}\left[r(X) \mid S = 1\right] &= \int \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)} \mathrm{d}P_1(x) \\
&= \int \left[\left(\frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)}\right)^2 - 1\right] \mathrm{d}P_0(x) + 1 \\
&= \mathrm{D}_{\chi^2}(P_1\|P_0) + 1.
\end{aligned} \tag{38}$$

Combining (36), (37), and (38) together, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[(h^*(X) - y_1(X))^2 \mid S = 1\right] \\
&\geq \left(\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right] - \sqrt{\epsilon \mathbb{E}\left[r(X) \mid S = 1\right]}\right)^2 \\
&= (A - B\sqrt{\epsilon})^2,
\end{aligned}$$

where A and B are defined in (30). □

Now we are in a position to prove the lower bound for $\epsilon_{\text{split},2}$.

Proof. By Lemma 3, for any classifier $h : \mathcal{X} \to [0,1]$ and

$$0 \le \epsilon < \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right],$$

if $\mathbb{E}\left[(h(X) - y_0(X))^2 \mid S = 0\right] \le \epsilon$ holds, then we have $\mathbb{E}\left[(h(X) - y_1(X))^2 \mid S = 1\right] \ge (A - B\sqrt{\epsilon})^2$, where

$$A = \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right], \; B = \sqrt{\mathsf{D}_{\chi^2}(P_1 \| P_0) + 1}.$$

Now we take $\epsilon = (A/(B+1))^2$. This ϵ satisfies our assumption since

$$\begin{aligned}\epsilon < \left(\frac{A}{B}\right)^2 &= \left(\frac{\int |y_1(x) - y_0(x)| \mathrm{d}P_1(x)}{\sqrt{\int \frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)} \mathrm{d}P_1(x)}}\right)^2 \\ &\le \int (y_1(x) - y_0(x))^2 \mathrm{d}P_0(x) \\ &= \mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = 0\right],\end{aligned}$$

where the second inequality is due to the Cauchy–Schwarz inequality. By Lemma 3, if

$$\mathbb{E}\left[(h(X) - y_0(X))^2 \mid S = 0\right] \le \left(\frac{A}{B+1}\right)^2,$$

then

$$\mathbb{E}\left[(h(X) - y_1(X))^2 \mid S = 1\right] \ge \left(\frac{A}{B+1}\right)^2.$$

Consequently, for any $h : \mathcal{X} \to [0,1]$,

$$\begin{aligned}&\max_{s \in \{0,1\}} \mathbb{E}\left[(h(X) - y_s(X))^2 \mid S = s\right] \\ &\ge \left(\frac{\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 1\right]}{\sqrt{\mathsf{D}_{\chi^2}(P_1 \| P_0) + 1} + 1}\right)^2.\end{aligned}$$

By symmetry, one can swap the positions of $S = 0$ and $S = 1$ and obtain

$$\begin{aligned}&\max_{s \in \{0,1\}} \mathbb{E}\left[(h(X) - y_s(X))^2 \mid S = s\right] \\ &\ge \max_{s \in \{0,1\}} \left\{\left(\frac{\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right]}{\sqrt{\mathsf{D}_{\chi^2}(P_s \| P_{1-s}) + 1} + 1}\right)^2\right\}.\end{aligned}$$

□

C. Extension to Cross Entropy Loss

As discussed in Remark 1, one can define the benefit-of-splitting under the cross entropy loss. We provide upper and lower bounds of the corresponding benefit-of-splitting as an extension of Theorem 1.

Proposition 4. *The cross-entropy-benefit-of-splitting* $\epsilon_{\textit{split},\mathsf{H}}$ *can be upper bounded by the minimum of*

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

and lower bounded by

$$\begin{aligned}&\mathsf{JS}(P_{X,Y|S=0} \| P_{X,Y|S=1}) - \mathsf{JS}(P_0 \| P_1) \\ &- \frac{1}{2}\left|\mathbb{E}\left[\mathsf{H}(y_0(X)) \mid S = 0\right] - \mathbb{E}\left[\mathsf{H}(y_1(X)) \mid S = 1\right]\right|.\end{aligned}$$

Proof. Recall that $\mathsf{H}(p,q) = \mathsf{D}_{\mathsf{KL}}(p \| q) + \mathsf{H}(p)$. Hence,

$$\begin{aligned}&\mathbb{E}\left[\mathsf{H}(y_s(X), h(X)) \mid S = s\right] \\ &= \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}(y_s(X) \| h(X)) \mid S = s\right] + \mathbb{E}\left[H(y_s(X)) \mid S = s\right],\end{aligned}$$

which leads to

$$\begin{aligned}&\max_{s \in \{0,1\}} \inf_{h : \mathcal{X} \to [0,1]} \mathbb{E}\left[\mathsf{H}(y_s(X), h(X)) \mid S = s\right] \\ &= \max_{s \in \{0,1\}} \mathbb{E}\left[H(y_s(X)) \mid S = s\right].\end{aligned} \tag{39}$$

Since $\max\{a_i + b_i\} \le \max\{a_i\} + \max\{b_i\}$,

$$\begin{aligned}&\inf_{h : \mathcal{X} \to [0,1]} \max_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{H}(y_s(X), h(X)) \mid S = s\right] \\ &\le \inf_{h : \mathcal{X} \to [0,1]} \max_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{KL}(y_s(X) \| h(X)) \mid S = s\right] \\ &\quad + \max_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{H}(y_s(X)) \mid S = s\right].\end{aligned} \tag{40}$$

Combining (39) and (40) implies $\epsilon_{\text{split},\mathsf{H}} \le \epsilon_{\text{split},\mathsf{KL}}$. Therefore, the upper bound of $\epsilon_{\text{split},\mathsf{KL}}$ in Theorem 1 is an upper bound of $\epsilon_{\text{split},\mathsf{H}}$ as well. Now we proceed to prove the lower bound.

$$\begin{aligned}&\inf_{h : \mathcal{X} \to [0,1]} \max_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{H}(y_s(X), h(X)) \mid S = s\right] \\ &\ge \frac{1}{2} \inf_{h : \mathcal{X} \to [0,1]} (\mathbb{E}\left[\mathsf{H}(y_0(X), h(X)) \mid S = 0\right] \\ &\qquad + \mathbb{E}\left[\mathsf{H}(y_1(X), h(X)) \mid S = 1\right]) \\ &= \frac{1}{2} \inf_{h : \mathcal{X} \to [0,1]} \sum_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}(y_s(X) \| h(X)) \mid S = s\right] \\ &\quad + \frac{1}{2} \sum_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{H}(y_s(X)) \mid S = s\right].\end{aligned} \tag{41}$$

By the proof of Lemma 2, we have

$$\begin{aligned}&\frac{1}{2} \inf_{h : \mathcal{X} \to [0,1]} \sum_{s \in \{0,1\}} \mathbb{E}\left[\mathsf{D}_{\mathsf{KL}}(y_s(X) \| h(X)) \mid S = s\right] \\ &= \mathsf{JS}(P_{X,Y|S=0} \| P_{X,Y|S=1}) - \mathsf{JS}(P_0 \| P_1).\end{aligned} \tag{42}$$

Combining (39), (41), (42) gives the desired lower bound. □

D. Proof of Proposition 1

Proof. First, note that

$$\inf_{\substack{h_s : \mathcal{X} \to [0,1] \\ \text{for } s \in \{0,1\}}} \mathbb{E}\left[|h_S(X) - y_S(X)|\right] = 0$$

as one can choose $h_s(x) = y_s(x)$. Hence, $\epsilon_{\text{split,pop}}$ is equal to

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]} \mathbb{E}\left[|h(X) - y_S(X)|\right]\\ &= \inf_{h:\mathcal{X}\to[0,1]} \sum_{s\in\{0,1\}} \Pr(S=s)\int |h(x)-y_s(x)|\mathrm{d}P_s(x).\end{aligned} \tag{43}$$

By the proof of Lemma 2, the optimal classifier of (43) is

$$h^*(x) = \begin{cases} y_0(x) & \text{if } \frac{\Pr(S=0)\cdot \mathrm{d}P_0(x)}{\Pr(S=1)\cdot \mathrm{d}P_1(x)} \geq 1\\ y_1(x) & \text{otherwise.}\end{cases}$$

By plugging the optimal classifier into (43), we can write $\epsilon_{\text{split,pop}}$ equivalently as

$$\begin{aligned}&\Pr(S=0)\cdot\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right]\\ &-\int |y_1(x)-y_0(x)|\big(\Pr(S=0)\cdot \mathrm{d}P_0(x)\\ &\qquad\qquad -\Pr(S=1)\cdot \mathrm{d}P_1(x)\big)_+.\end{aligned}$$

The desired upper bound can be obtained by dropping the negative term. Now we proceed to prove the lower bound. Since

$$\begin{aligned}&\int |y_1(x)-y_0(x)|(\Pr(S=0)\mathrm{d}P_0(x) - \Pr(S=1)\mathrm{d}P_1(x))_+\\ &\leq \Pr(S=0)\int \left(\frac{\mathrm{d}P_0(x)}{\mathrm{d}P_1(x)} - \frac{\Pr(S=1)}{\Pr(S=0)}\right)_+ \mathrm{d}P_1(x)\\ &= \Pr(S=0)\cdot E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1),\end{aligned}$$

where $E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)$ is the E_γ-divergence, we have

$$\begin{aligned}\epsilon_{\text{split,pop}} \geq \Pr(S=0)\Big(&\mathbb{E}\left[|y_1(X)-y_0(X)| \mid S=0\right]\\ &- E_{\frac{\Pr(S=1)}{\Pr(S=0)}}(P_0\|P_1)\Big).\end{aligned}$$

□

APPENDIX C
PROOFS FOR SECTION IV

A. Proof of Theorem 2

Recall that the false error rate is the maximum between false positive rate and false negative rate

$$\begin{aligned}\mathsf{FER}_s(h) \triangleq \max\{&\mathbb{E}\left[h(X) \mid Y=0, S=s\right],\\ &\mathbb{E}\left[1-h(X) \mid Y=1, S=s\right]\}.\end{aligned}$$

We prove the following lemma which will be used in the proof of Theorem 2.

Lemma 4. *The false error rate has the following equivalent expressions*

$$\begin{aligned}\mathsf{FER}_s(h) &= \max\left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X)) \mid S=s\right]}{\Pr(Y=0\mid S=s)},\right.\\ &\qquad \left.1-\frac{\mathbb{E}\left[h(X)y_s(X)\mid S=s\right]}{\Pr(Y=1\mid S=s)}\right\}\\ &= \max\left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0\mid S=s)},\right.\\ &\qquad \left.1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1\mid S=s)}\right\}.\end{aligned}$$

where $f_s(x) \triangleq \frac{\Pr(S=s|X=x)}{\Pr(S=s)}$.

Proof. The proof follows directly from Bayes's rule,

$$\begin{aligned}\mathrm{d}P_{X|Y=0,S=s} &= \frac{1-y_s(x)}{\Pr(Y=0\mid S=s)}\mathrm{d}P_{X|S=s}\\ &= \frac{1-y_s(x)}{\Pr(Y=0\mid S=s)}\cdot f_s(x)\mathrm{d}P_X.\end{aligned}$$

□

Now we are in a position to prove Theorem 2.

Proof. By Lemma 4, $\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\mathsf{FER}_s(h)$ can be equivalently written as

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]}\max_{s\in\{0,1\}}\left\{\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0\mid S=s)},\right.\\ &\qquad\qquad \left.1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1\mid S=s)}\right\}\\ &= \inf_{h:\mathcal{X}\to[0,1]}\max_{\boldsymbol{\mu}\in\Delta_4} G(\boldsymbol{\mu},h),\end{aligned} \tag{44}$$

where $\boldsymbol{\mu} \triangleq (\mu_{0,0},\mu_{0,1},\mu_{1,0},\mu_{1,1})$ and $G(\boldsymbol{\mu},h)$ is defined as

$$\begin{aligned}&\sum_{s\in\{0,1\}}\Big(\mu_{s,0}\frac{\mathbb{E}\left[h(X)(1-y_s(X))f_s(X)\right]}{\Pr(Y=0\mid S=s)}\\ &\qquad +\mu_{s,1}\Big(1-\frac{\mathbb{E}\left[h(X)y_s(X)f_s(X)\right]}{\Pr(Y=1\mid S=s)}\Big)\Big)\\ &= \sum_{s\in\{0,1\}}\mu_{s,1} + \mathbb{E}\Big[\sum_{s\in\{0,1\}}\Big(\frac{\mu_{s,0}(1-y_s(X))f_s(X)}{\Pr(Y=0\mid S=s)}\\ &\qquad\qquad -\frac{\mu_{s,1}y_s(X)f_s(X)}{\Pr(Y=1\mid S=s)}\Big)h(X)\Big].\end{aligned}$$

By denoting

$$\begin{aligned}\phi_{s,0}(x) &\triangleq \frac{(1-y_s(x))f_s(x)}{\Pr(Y=0\mid S=s)}\\ \phi_{s,1}(x) &\triangleq \frac{-y_s(x)f_s(x)}{\Pr(Y=1\mid S=s)},\end{aligned}$$

we can write

$$G(\boldsymbol{\mu},h) = \sum_{s\in\{0,1\}}\mu_{s,1} + \mathbb{E}\left[\sum_{s,i\in\{0,1\}}\mu_{s,i}\phi_{s,i}(X)h(X)\right].$$

We next use Ky Fan's min-max theorem [24] (see Lemma 1) to swap the positions of infimum and maximum. First, Δ_4 is a compact set and for any $h:\mathcal{X}\to[0,1]$, $G(\cdot,h)$ is continuous on Δ_4. Furthermore, for any $h:\mathcal{X}\to[0,1]$, $G(\cdot,h)$ is linear over Δ_4; for any $\boldsymbol{\mu}\in\Delta_4$, $G(\boldsymbol{\mu},\cdot)$ is convex-like over all (measurable) classifiers from $\mathcal{X}$ to $[0,1]$. Hence, we have

$$\inf_{h:\mathcal{X}\to[0,1]}\max_{\boldsymbol{\mu}\in\Delta_4} G(\boldsymbol{\mu},h) = \max_{\boldsymbol{\mu}\in\Delta_4}\inf_{h:\mathcal{X}\to[0,1]} G(\boldsymbol{\mu},h). \tag{45}$$

Next, we prove that, for any fixed $\boldsymbol{\mu}\in\Delta_4$,

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]}\mathbb{E}\left[\sum_{s,i\in\{0,1\}}\mu_{s,i}\phi_{s,i}(X)h(X)\right]\\ &= \mathbb{E}\left[\inf_{h:\mathcal{X}\to[0,1]}\sum_{s,i\in\{0,1\}}\mu_{s,i}\phi_{s,i}(X)h(X)\right].\end{aligned} \tag{46}$$

One direction LHS $\geq$ RHS can be obtained directly since for any $h : \mathcal{X} \to [0,1]$

$$\mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right] \geq \mathbb{E}\left[\inf_{h:\mathcal{X}\to[0,1]} \sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h(X)\right].$$

Note that the infimum in the RHS of (46) is point-wise. For any fixed $x \in \mathcal{X}$, the following optimization problem

$$\inf_{\bar{h}\in[0,1]} \sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x)\bar{h}$$

has an optimal solution $\bar{h}^* = \mathbb{I}[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x) \leq 0]$. Hence, there is a measurable classifier which can achieve the point-wise infimum inside the expectation of the RHS in (46): $h^*(x) = \mathbb{I}[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x) \leq 0]$. Consequently, the RHS of (46) can be simplified as

$$\text{RHS} = \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)\right)_-\right], \tag{47}$$

where for $a \in \mathbb{R}$, $(a)_- \triangleq \min\{a, 0\}$. Since the LHS of (46) is an infimum over all measurable classifiers, using the classifier h^* leads to

$$\begin{aligned}\text{LHS} &\leq \mathbb{E}\left[\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)h^*(X)\right] \\ &= \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)\right)_-\right] = \text{RHS}.\end{aligned}$$

Combining (44–47) together implies

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathsf{FER}_s(h) \\ &= \max_{\boldsymbol{\mu}\in\Delta_4} \left\{\sum_{s\in\{0,1\}} \mu_{s,1} + \mathbb{E}\left[\left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(X)\right)_-\right]\right\}.\end{aligned}$$

Similarly, one can prove that

$$\begin{aligned}&\max_{s\in\{0,1\}} \inf_{h:\mathcal{X}\to[0,1]} \mathsf{FER}_s(h) \\ &= \max_{s\in\{0,1\}} \max_{\boldsymbol{\nu}^{(s)}\in\Delta_2} \left\{\nu_1^{(s)} + \mathbb{E}\left[\left(\sum_{i\in\{0,1\}} \nu_i^{(s)}\phi_{s,i}(X)\right)_-\right]\right\}.\end{aligned}$$

□

B. Proof of Proposition 2

We start with a useful lemma which will be used in the proof of Proposition 2.

Lemma 5. *Let $f : \mathcal{X} \times \mathbb{R}^k \to \mathbb{R}$ be a bounded measurable function. For a fixed $x \in \mathcal{X}$, if $v(x, w_0) \in \mathbb{R}^k$ is a supergradient of $f(x,\cdot)$ at w_0:*

$$f(x,w) - f(x,w_0) \leq v(x,w_0)^T(w - w_0), \tag{48}$$

then $\mathbb{E}[v(X, w_0)]$ is a supergradient of $\mathbb{E}[f(X,\cdot)]$ at w_0:

$$\mathbb{E}[f(X,w)] - \mathbb{E}[f(X,w_0)] \leq \mathbb{E}[v(X,w_0)]^T(w - w_0).$$

The proof of Lemma 5 follows directly by taking expectation on both sides of (48). We refer the readers to [111] for a more general result on the interchangeability of subdifferentiation and (conditional) expectation. Now we are in a position to prove Proposition 2.

Proof. Consider a function

$$g(x,\boldsymbol{\mu}) \triangleq \sum_{s\in\{0,1\}} \mu_{s,1} + \left(\sum_{s,i\in\{0,1\}} \mu_{s,i}\phi_{s,i}(x)\right)_-.$$

For a fixed x, $g(x,\cdot)$ has a supergradient at $\boldsymbol{\mu} \triangleq (\mu_{0,0}, \mu_{0,1}, \mu_{1,0}, \mu_{1,1})$:

$$\left(i + \phi_{s,i}(x)\mathbb{I}\Big[\sum_{s',i'} \mu_{s',i'}\phi_{s',i'}(x) < 0\Big]\right).$$

Therefore, by Lemma 5, g has a supergradient at $\boldsymbol{\mu}$:

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

Now we introduce auxiliary functions

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

By Bayes's rule and the definition of $\phi_{s,i}$ (see (5)), we have

$$\psi_{s,i}(x)\cdot \mathrm{d}P_{X|S=s}(x) = \phi_{s,i}(x)\cdot \mathrm{d}P_X(x).$$

Hence, the supergradient of g given above can be rewritten as

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

Similarly, one can obtain a closed-form supergradient of g_s. □

APPENDIX D
PROOFS FOR SECTION V

A. Proof of Theorem 3

We first recall a useful lemma which can be proved by the variational representation [112] of total variation distance.

Lemma 6. *For any measurable and non-negative function $f : \mathcal{X} \to \mathbb{R}^+$,*

$$|\mathbb{E}[f(X_0)] - \mathbb{E}[f(X_1)]| \leq \|f\|_\infty \mathsf{D}_{\mathsf{TV}}(P_0\|P_1),$$

where $X_0 \sim P_0$ and $X_1 \sim P_1$.

Now we are in a position to prove Theorem 3.

Proof. We prove the upper bound first. Let h_s^* be an optimal classifier for the group $s \in \{0,1\}$, i.e., $h_s^* \in \operatorname{argmin}_{h\in\mathcal{H}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]$. Then

$$\begin{aligned}&\inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &\le \max\{\mathbb{E}\left[|h_0^*(X) - y_1(X)| \mid S = 1\right],\\ &\qquad \mathbb{E}\left[|h_0^*(X) - y_0(X)| \mid S = 0\right]\}.\end{aligned}$$

By the triangle inequality, $\mathbb{E}\left[|h_0^*(X) - y_1(X)| \mid S = 1\right]$ can be upper bounded by

$$\mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = 1\right] + \mathbb{E}\left[|h_1^*(X) - y_1(X)| \mid S = 1\right].$$

Therefore,

$$\begin{aligned}&\inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &\le \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = 1\right]\\ &\quad + \max_{s\in\{0,1\}} \mathbb{E}\left[|h_s^*(X) - y_s(X)| \mid S = s\right],\end{aligned}$$

which implies that

$$\epsilon_{\text{split}}^{\mathcal{H}} \le \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = 1\right].$$

By symmetry, we obtain the desired upper bound for $\epsilon_{\text{split}}^{\mathcal{H}}$. Now we proceed to prove the lower bound for $\epsilon_{\text{split}}^{\mathcal{H}}$. By the triangle inequality, $\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 0\right]$ can be lower bounded by

$$\begin{aligned}&\mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = 0\right] - \mathbb{E}\left[|h_0^*(X) - y_0(X)| \mid S = 0\right]\\ &- \mathbb{E}\left[|h_1^*(X) - y_1(X)| \mid S = 0\right].\end{aligned}$$

By Lemma 6,

$$\begin{aligned}&\mathbb{E}\left[|h_1^*(X) - y_1(X)| \mid S = 0\right]\\ &\le \mathbb{E}\left[|h_1^*(X) - y_1(X)| \mid S = 1\right] + \mathsf{D}_{\mathsf{TV}}(P_0\|P_1).\end{aligned}$$

Therefore,

$$\begin{aligned}&\mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = 0\right]\\ &\ge \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = 0\right] - 2\lambda - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1).\end{aligned}$$

where $\lambda \triangleq \sum_{s\in\{0,1\}} \mathbb{E}\left[|h_s^*(X) - y_s(X)| \mid S = s\right]/2$. Hence,

$$\begin{aligned}&\max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right]\\ &\ge \max_{s\in\{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right] - 2\lambda - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1).\end{aligned} \tag{49}$$

By a slight modification of the proof of Theorem 1, we have

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &\ge \frac{1}{2}\left(\max_{s\in\{0,1\}} \mathbb{E}\left[|y_1(X) - y_0(X)| \mid S = s\right] - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1)\right).\end{aligned} \tag{50}$$

Substituting (49) into (50) leads to

$$\begin{aligned}&\inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &\ge \inf_{h:\mathcal{X}\to[0,1]} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &\ge \frac{1}{2}\Big(\max_{s\in\{0,1\}} \mathbb{E}\left[|h_1^*(X) - h_0^*(X)| \mid S = s\right]\\ &\qquad - 2\lambda - 2\mathsf{D}_{\mathsf{TV}}(P_0\|P_1)\Big).\end{aligned} \tag{51}$$

Finally, since $\max\{a,b\} \le a + b$ and $\{h_s^*\}_{s\in\{0,1\}}$ is the set of optimal split classifiers, then

$$\begin{aligned}&\max_{s\in\{0,1\}} \inf_{h\in\mathcal{H}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\\ &= \max_{s\in\{0,1\}} \mathbb{E}\left[|h_s^*(X) - y_s(X)| \mid S = s\right] \le 2\lambda.\end{aligned} \tag{52}$$

Combining (51) with (52) gives the desired lower bound. □

B. Proof of Proposition 3

Proof. By the triangle inequality, we can upper bound $\inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]$ by I + II where

$$\begin{aligned}\mathrm{I} &\triangleq \inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \mathbb{E}\left[|h(X) - h^*(X)| \mid S = s\right],\\ \mathrm{II} &\triangleq \max_{s\in\{0,1\}} \mathbb{E}\left[|h^*(X) - y_s(X)| \mid S = s\right],\end{aligned}$$

and h^* is defined in (22). Since $\max\{a,b\} \le a + b$, $\mathrm{I} \le 2\inf_{h\in\mathcal{H}} \mathbb{E}\left[\left|h(\bar{X}) - h^*(\bar{X})\right|\right]$ where the random variable $\bar{X}$ follows the probability distribution $(P_0 + P_1)/2$. By Barron's approximation bounds [97],

$$\inf_{h\in\mathcal{H}} \mathbb{E}\left[\left|h(\bar{X}) - h^*(\bar{X})\right|\right] \le \frac{\mathsf{diam}(\mathcal{X})C}{\sqrt{k}}, \tag{53}$$

where the constant $C \triangleq \int_{\mathbb{R}^d} \|w\|_2 |\widetilde{h^*}(w)| \mathrm{d}w$ and $\widetilde{h^*}(w) \triangleq \frac{1}{(2\pi)^d}\int_{\mathcal{X}} h^*(x)\exp(-iwx)\mathrm{d}x$. Moreover, by the proof of Theorem 1 (see Appendix B-B), we have

$$\begin{aligned}\mathrm{II} &\le \min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]}\\ &\quad \times \sqrt{1 - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1)}.\end{aligned}$$

To summarize, if the hypothesis class contains feedforward neural network models with one layer of sigmoidal functions, the $\mathcal{H}$-benefit-of-splitting $\epsilon_{\text{split}}^{\mathcal{H}}$ can be upper bounded by

$$\begin{aligned}&\min_{s\in\{0,1\}} \sqrt{\mathbb{E}\left[(y_1(X) - y_0(X))^2 \mid S = s\right]} \cdot \sqrt{1 - \mathsf{D}_{\mathsf{TV}}(P_0\|P_1)}\\ &+ \frac{2\mathsf{diam}(\mathcal{X})C}{\sqrt{k}}.\end{aligned}$$

□

C. Proof of Corollary 1

We approach Corollary 1 by proving a more general result.

Lemma 7. *For any hypothesis $\mathcal{H}$, there exists a probability distribution $Q_{S,X,Y}$ whose $\mathcal{H}$-benefit-of-splitting is at least*

$$\frac{1}{2}\sup_{\substack{h_1,h_0\in\mathcal{H}\\ x\in\mathcal{X}}} |h_1(x) - h_0(x)|.$$

Proof. For any $\epsilon > 0$, there exist two classifiers $h_1^*, h_0^* \in \mathcal{H}$ and $x^* \in \mathcal{X}$ such that

$$|h_1^*(x^*) - h_0^*(x^*)| \ge \sup_{\substack{h_1,h_0\in\mathcal{H}\\ x\in\mathcal{X}}} |h_1(x) - h_0(x)| - \epsilon. \tag{54}$$

Now we construct a probability distribution $Q_{S,X,Y}$ with $Q_{Y|X,S}(1|x,s) = h_s^*(x)$, $Q_{X|S=s}(x) = \delta(x - x^*)$, $Q_S(s) = 0.5$ for $s \in \{0,1\}$ where $\delta(\cdot)$ is the Dirac delta function. Our

lower bound in Theorem 3 implies that $\epsilon^{\mathcal{H}}_{\text{split}} \geq \frac{1}{2}|h_1^*(x^*) - h_0^*(x^*)|$ which, due to (54), can be further lower bounded by $\frac{1}{2}(\sup_{h_1,h_0\in\mathcal{H},x\in\mathcal{X}}|h_1(x)-h_0(x)|-\epsilon)$. Since this lower bound of $\epsilon^{\mathcal{H}}_{\text{split}}$ holds for any $\epsilon > 0$, one can let ϵ be sufficiently small which leads to the desired conclusion. □

D. Proof of Theorem 4

We introduce the empirical benefit-of-splitting and bound its difference from the sample-limited-splitting (see Definition 7).

Definition 8. For a given hypothesis class $\mathcal{H}$ and n_s i.i.d. samples $\{(x_{s,i}, y_{s,i})\}_{i=1}^{n_s}$ from group $s \in \{0,1\}$, the empirical-splitting is defined as

$$\begin{aligned}\hat{\epsilon}_{\text{split,emp}} \triangleq \inf_{h\in\mathcal{H}} \max_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s}|h(x_{s,i}) - y_{s,i}|}{n_s} \\ - \max_{s\in\{0,1\}} \inf_{h\in\mathcal{H}} \frac{\sum_{i=1}^{n_s}|h(x_{s,i}) - y_{s,i}|}{n_s}.\end{aligned} \tag{55}$$

Lemma 8. *Let $\mathcal{H}$ be a hypothesis class from $\mathcal{X}$ to $\{0,1\}$ with VC dimension D. Then with probability at least $1-\delta$,*

$$|\hat{\epsilon}_{split} - \hat{\epsilon}_{split,emp}| \leq 4 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(16/\delta)}{n_s}},$$

where n_s is the number of samples from group $s \in \{0,1\}$.

Proof. Corollary 3.8 and Theorem 4.3 in [98] together imply that with probability at least $1-\delta$, for any $s \in \{0,1\}$ and $h \in \mathcal{H}$,

$$\begin{aligned}&\left|\frac{\sum_{i=1}^{n_s}|h(x_{s,i}) - y_{s,i}|}{n_s} - \mathbb{E}\left[|h(X) - y_s(X)| \mid S = s\right]\right| \\ &\leq 2\sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.\end{aligned}$$

Therefore, for any $h_s \in \mathcal{H}$ with $s \in \{0,1\}$

$$\begin{aligned}&\Big|\max_{s\in\{0,1\}} \frac{\sum_{i=1}^{n_s}|h_s(x_{s,i}) - y_{s,i}|}{n_s} \\ &- \max_{s\in\{0,1\}} \mathbb{E}\left[|h_s(X) - y_s(X)| \mid S = s\right]\Big| \\ &\leq 2 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.\end{aligned} \tag{56}$$

Recall that

$$\begin{aligned}\hat{\epsilon}_{\text{split}} = \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}^*(X) - y_s(X)| \mid S = s\right] \\ - \max_{s\in\{0,1\}} \mathbb{E}\left[|\hat{h}_s^*(X) - y_s(X)| \mid S = s\right].\end{aligned}$$

Now by (56), we conclude that

$$|\hat{\epsilon}_{\text{split}} - \hat{\epsilon}_{\text{split,emp}}| \leq 4 \max_{s\in\{0,1\}} \sqrt{\frac{2D\log(6n_s) + 2\log(8/\delta)}{n_s}}.$$

□

Since the upper and lower bounds of $\epsilon^{\mathcal{H}}_{\text{split}}$ (see Theorem 3) hold for any underlying distribution $P_{S,X,Y}$. One can plug in the empirical distribution and obtain the corresponding bounds for $\hat{\epsilon}_{\text{split,emp}}$. Then we obtain the desired bounds for $\hat{\epsilon}_{\text{split}}$ by using Lemma 8 for bounding the difference between $\hat{\epsilon}_{\text{split,emp}}$ and $\hat{\epsilon}_{\text{split}}$.

APPENDIX E
SUPPORTING RESULTS FOR EXPERIMENTS

A. Closed-form Expression of $\epsilon_{split,\mathsf{FER}}$

Proof. For the distributions we construct, one can choose $\{y_s(x)\}_{s\in\{0,1\}}$ as the split classifiers which lead to zero false error rate. Therefore, the problem remains computing the false error rate of the optimal group-blind classifier. First, the labeling functions naturally divide $\mathbb{R}^2$ into four parts: $\text{I} \triangleq \{x \mid y_0(x) = 1, y_1(x) = 1\}$, $\text{II} \triangleq \{x \mid y_0(x) = 1, y_1(x) = 0\}$, $\text{III} \triangleq \{x \mid y_0(x) = 0, y_1(x) = 0\}$, $\text{IV} \triangleq \{x \mid y_0(x) = 0, y_1(x) = 1\}$. Clearly, in order to be an optimal group-blind classifier, h must satisfy $h(x) = 1$ on I and $h(x) = 0$ on III. We define $h|_{\text{II}}$ and $h|_{\text{IV}}$ as

$$h|_{\text{II}}(x) \triangleq \begin{cases} h(x) & \text{if } x \in \text{II} \\ 0 & \text{otherwise,}\end{cases}$$

$$h|_{\text{IV}}(x) \triangleq \begin{cases} h(x) & \text{if } x \in \text{IV} \\ 0 & \text{otherwise.}\end{cases}$$

Due to our construction of the labeling functions and Lemma 4, $\epsilon_{\text{split,FER}}$ is equal to

$$\begin{aligned}&\inf_{h:\mathcal{X}\to[0,1]} \max\Big\{\frac{\mathbb{E}\left[h(X)(1-y_0(X)) \mid S=0\right]}{\Pr(Y=0 \mid S=0)}, \\ &\qquad 1 - \frac{\mathbb{E}\left[h(X)y_0(X) \mid S=0\right]}{\Pr(Y=1 \mid S=0)}, \\ &\qquad \frac{\mathbb{E}\left[h(X)(1-y_1(X)) \mid S=1\right]}{\Pr(Y=0 \mid S=1)}, \\ &\qquad 1 - \frac{\mathbb{E}\left[h(X)y_1(X) \mid S=1\right]}{\Pr(Y=1 \mid S=1)}\Big\} \\ &= \inf_{h:\mathcal{X}\to[0,1]} \max\{2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=0\right], \\ &\qquad 1 - 2\Pr(X \in \text{I} \mid S=0) - 2\mathbb{E}\left[h|_{\text{II}}(X) \mid S=0\right], \\ &\qquad 2\mathbb{E}\left[h|_{\text{II}}(X) \mid S=1\right], \\ &\qquad 1 - 2\Pr(X \in \text{I} \mid S=1) - 2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=1\right]\},\end{aligned}$$

which is equivalent to

$$\begin{aligned}&\max\Big\{\inf_{h|_{\text{II}}:\mathcal{X}\to[0,1]} \max\{2\mathbb{E}\left[h|_{\text{II}}(X) \mid S=1\right], \\ &\qquad 1 - 2\Pr(X \in \text{I} \mid S=0) - 2\mathbb{E}\left[h|_{\text{II}}(X) \mid S=0\right]\}, \\ &\qquad \inf_{h|_{\text{IV}}:\mathcal{X}\to[0,1]} \max\{2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=0\right], \\ &\qquad 1 - 2\Pr(X \in \text{I} \mid S=1) - 2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=1\right]\}\Big\} \\ &= \inf_{h|_{\text{IV}}:\mathcal{X}\to[0,1]} \max\{2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=0\right], \\ &\qquad 1 - 2\Pr(X \in \text{I} \mid S=1) - 2\mathbb{E}\left[h|_{\text{IV}}(X) \mid S=1\right]\},\end{aligned} \tag{57}$$

where the last step is because of symmetry. Since $2\max\{a,b\} \geq a+b$, $\epsilon_{\text{split,FER}}$ can be lower bounded by

$$\begin{aligned}
&\frac{1}{2} - \Pr(X \in \mathrm{I} \mid S=1) \\
&+ \inf_{h|_{\mathrm{IV}}:\mathcal{X}\to[0,1]} \mathbb{E}\left[h|_{\mathrm{IV}}(X) \mid S=0\right] - \mathbb{E}\left[h|_{\mathrm{IV}}(X) \mid S=1\right] \\
&= \frac{1}{2} - \Pr(X \in \mathrm{I} \mid S=1) \\
&\quad + \inf_{h|_{\mathrm{IV}}:\mathcal{X}\to[0,1]} \int_{\mathrm{IV}} h|_{\mathrm{IV}}(x)(\mathrm{d}P_0(x) - \mathrm{d}P_1(x)) \\
&= \frac{1}{2} - \Pr(X \in \mathrm{I} \mid S=1) - \int_{\mathrm{IV}} (\mathrm{d}P_1(x) - \mathrm{d}P_0(x))_+.
\end{aligned}$$

Since $P_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_0)$ and $P_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_1)$, by comparing their probability density functions, we have

$$\begin{aligned}
\mathcal{A} &\triangleq \{x \in \mathrm{IV} \mid \mathrm{d}P_1(x) > \mathrm{d}P_0(x)\} = \{x \in \mathrm{IV} \mid x_2 < 0\} \\
&= \{x \in \mathbb{R}^2 \mid y_1(x) = 1, x_2 < 0\}.
\end{aligned}$$

Therefore, $\epsilon_{\text{split,FER}}$ can be lower bounded by

$$\begin{aligned}
\frac{1}{2}\Big(&1 - 2\Pr(X \in \mathrm{I} \mid S=1) - 2\Pr(X \in \mathcal{A} \mid S=1) \\
&+ 2\Pr(X \in \mathcal{A} \mid S=0)\Big).
\end{aligned} \tag{58}$$

By symmetry, we have

$$\begin{aligned}
\Pr(X \in \mathrm{I} \mid S=1) &= \Pr(X \in \mathrm{III} \mid S=1), \\
\Pr(X \in \mathrm{II} \mid S=1) &= \Pr(X \in \mathrm{IV} \mid S=1),
\end{aligned}$$

which leads to

$$\begin{aligned}
&1 - 2\Pr(X \in \mathrm{I} \mid S=1) - 2\Pr(X \in \mathcal{A} \mid S=1) \\
&= 2\Pr(X \in \mathrm{IV} \backslash \mathcal{A} \mid S=1) \\
&= 2\Pr(X \in \mathcal{A} \mid S=0),
\end{aligned} \tag{59}$$

where the last step is by symmetry again. Therefore, the lower bound of $\epsilon_{\text{split,FER}}$ in (58) is $2\Pr(X \in \mathcal{A} \mid S=0)$. On the other hand, we can design a classifier $h|^*_{\mathrm{IV}}(x) = 1$ if $x \in \mathcal{A}$; $h|^*_{\mathrm{IV}}(x) = 0$ otherwise. By (57), $\epsilon_{\text{split,FER}}$ can be upper bounded by

$$\begin{aligned}
&\max\{2\mathbb{E}\left[h|^*_{\mathrm{IV}}(X) \mid S=0\right], \\
&\qquad 1 - 2\Pr(X \in \mathrm{I} \mid S=1) - 2\mathbb{E}\left[h|^*_{\mathrm{IV}}(X) \mid S=1\right]\} \\
&= \max\{2\Pr(X \in \mathcal{A} \mid S=0), \\
&\qquad 1 - 2\Pr(X \in \mathrm{I} \mid S=1) - 2\Pr(X \in \mathcal{A} \mid S=1)\}.
\end{aligned}$$

The above upper bound is equal to $2\Pr(X \in \mathcal{A} \mid S=0)$ due to (59). Hence, $\epsilon_{\text{split,FER}} = 2\Pr(X \in \mathcal{A} \mid S=0)$. □

B. Total Variation Distance Estimation

We provide details on how we estimate the total variation distance $\mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)$ by using n_s i.i.d. unlabeled data $\{x_{s,i}\}_{i=1}^{n_s}$ drawn from each group $s \in \{0,1\}$. By applying Baye's rule, we can write the density ratio equivalently as

$$\begin{aligned}
\frac{\mathrm{d}P_1(x)}{\mathrm{d}P_0(x)} &= \frac{\mathrm{d}P_{X|S=1}(x)}{\mathrm{d}P_{X|S=0}(x)} \\
&= \frac{\Pr(S=1 \mid X=x)}{1 - \Pr(S=1 \mid X=x)} \cdot \frac{1 - \Pr(S=1)}{\Pr(S=1)},
\end{aligned}$$

which leads to an equivalent expression of $\mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)$:

$$\begin{aligned}
&\int \mathbb{I}\left[\frac{\Pr(S=1 \mid X=x)}{1 - \Pr(S=1 \mid X=x)} \geq \frac{\Pr(S=1)}{1 - \Pr(S=1)}\right] \mathrm{d}P_1(x) \\
&- \int \mathbb{I}\left[\frac{\Pr(S=1 \mid X=x)}{1 - \Pr(S=1 \mid X=x)} \geq \frac{\Pr(S=1)}{1 - \Pr(S=1)}\right] \mathrm{d}P_0(x).
\end{aligned} \tag{60}$$

This expression gives rise to the following procedure of estimating the total variation distance.

- Compute a constant $\alpha = \frac{n_1}{n_0+n_1}$ to estimate the marginal probability $\Pr(S=1)$ and train a classifier $s(x)$ to approximate the conditional distribution $\Pr(S=1 \mid X=x)$. In particular, we use a feed-forward neural network for $s(x)$, which consists of one hidden layer with 100 neurons and `ReLU` activation, and a soft-max readout layer. We adopt cross entropy as the loss function, set learning rate to be 0.001, and use `AdamOptimizer` [113] to train the datasets with batch size 200. To avoid overfitting, we hold 10% of the samples as a validation set, and terminate training once the validation loss is not improving by 10^{-4} for the next 10 consecutive epochs (i.e., early stopping), and the maximum number of epochs is set to be 200.
- By plugging α and $s(x)$ into (60) and using i.i.d. samples to estimate the integrals (i.e., expectations), we obtain the following approximation of $\mathrm{D}_{\mathsf{TV}}(P_0 \| P_1)$:

$$\begin{aligned}
&\frac{1}{n_1}\sum_{i=1}^{n_1} \mathbb{I}\left[\frac{s(x_{1,i})}{1 - s(x_{1,i})} \geq \frac{\alpha}{1-\alpha}\right] \\
&- \frac{1}{n_0}\sum_{i=1}^{n_0} \mathbb{I}\left[\frac{s(x_{0,i})}{1 - s(x_{0,i})} \geq \frac{\alpha}{1-\alpha}\right].
\end{aligned}$$

We remark that estimating information-theoretic measures has been studied in e.g., [26], [27], [101], [114], [115].

REFERENCES

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

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

[3] M. K. Cho, "Racial and ethnic categories in biomedical research: there is no baby in the bathwater," *J. Law Med. Ethics*, vol. 34, no. 3, pp. 497–499, 2006.

[4] J. N. Cohn, "The use of race and ethnicity in medicine: lessons from the african-american heart failure trial," *J. Law Med. Ethics*, vol. 34, no. 3, pp. 552–554, 2006.

[5] J. Perez-Rodriguez and A. de la Fuente, "Now is the time for a postracial medicine: Biomedical research, the national institutes of health, and the perpetuation of scientific racism," *Am. J. Bioeth.*, vol. 17, no. 9, pp. 36–47, 2017.

[6] 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

[7] 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.

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] ——, "On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning," in *Proc. 2018 IEEE Int. Symp. on Inf. Theory*, 2018, pp. 126–130.

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

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

[18] 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.

[19] 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.

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

[21] 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.

[22] J. Vanschoren, J. N. van Rijn, B. Bischl, and L. Torgo, "Openml: Networked science in machine learning," *SIGKDD Explorations*, vol. 15, no. 2, pp. 49–60, 2013.

[23] 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.

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

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

[26] 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.

[27] 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.

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

[29] 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.

[30] 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.

[31] 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.

[32] S. Li, A. Khisti, and A. Mahajan, "Information-theoretic privacy for smart metering systems with a rechargeable battery," *IEEE Trans. Inf. Theory*, vol. 64, no. 5, pp. 3679–3695, 2018.

[33] B. Rassouli and D. Gündüz, "Optimal utility-privacy trade-off with total variation distance as a privacy measure," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 594–603, 2019.

[34] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 11, pp. 1623–1636, 2009.

[35] Y. O. Basciftci, Y. Wang, and P. Ishwar, "On privacy-utility tradeoffs for constrained data release mechanisms," in *2016 Information Theory and Applications Workshop (ITA)*. IEEE, 2016, pp. 1–6.

[36] N. Takbiri, A. Houmansadr, D. L. Goeckel, and H. Pishro-Nik, "Privacy against statistical matching: Inter-user correlation," in *Proc. 2018 IEEE Int. Symp. on Inf. Theory*, 2018, pp. 1036–1040.

[37] A. Nageswaran and P. Narayan, "Data privacy for a ρ-recoverable function," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3470–3488, 2019.

[38] J. Liao, L. Sankar, F. P. Calmon, and V. Y. Tan, "Hypothesis testing under maximal leakage privacy constraints," in *Proc. 2017 IEEE Int. Symp. on Inf. Theory*, 2017, pp. 779–783.

[39] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *IEEE Inf. Theory Workshop (ITW)*, 2014, pp. 501–505.

[40] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," in *Proc. 37th Annu. Allerton Conf. Commun. Control Comput.*, 1999, pp. 368–377.

[41] H. Hsu, S. Asoodeh, S. Salamatian, and F. P. Calmon, "Generalizing bottleneck problems," in *Proc. 2018 IEEE Int. Symp. on Inf. Theory*, 2018, pp. 531–535.

[42] P. Kairouz, S. Oh, and P. Viswanath, "Extremal mechanisms for local differential privacy," in *Advances in Neural Information Processing Systems*, 2014, pp. 2879–2887.

[43] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *J. Amer. Statist. Assoc.*, vol. 60, no. 309, pp. 63–69, 1965.

[44] J. Geumlek and K. Chaudhuri, "Profile-based privacy for locally private computations," in *Proc. 2019 IEEE Int. Symp. on Inf. Theory*, 2019, pp. 537–541.

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

[46] F. P. Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. 50th Annu. Allerton Conf. Commun. Control Comput.*, 2012, pp. 1401–1408.

[47] M. Bertran, N. Martinez, A. Papadaki, Q. Qiu, M. Rodrigues, G. Reeves, and G. Sapiro, "Adversarially learned representations for information obfuscation and inference," in *Proc. 36th International Conference on Machine Learning*, 2019, pp. 614–623.

[48] H. Wang, L. Vo, F. P. Calmon, M. Médard, K. R. Duffy, and M. Varia, "Privacy with estimation guarantees," *IEEE Trans. Inf. Theory*, vol. 65, no. 12, pp. 8025–8042, 2019.

[49] H. Hsu, S. Asoodeh, and F. P. Calmon, "Information-theoretic privacy watchdogs," in *Proc. 2019 IEEE Int. Symp. on Inf. Theory*, 2019, pp. 552–556.

[50] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Trans. Inf. Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.

[51] M. Diaz, H. Wang, F. P. Calmon, and L. Sankar, "On the robustness of information-theoretic privacy measures and mechanisms," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 1949–1978, 2020.

[52] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[53] I. Kuzborskij and F. Orabona, "Stability and hypothesis transfer learning," in *Proc. 30th International Conference on Machine Learning*, 2013, pp. 942–950.

[54] S. Kpotufe and G. Martinet, "Marginal singularity, and the benefits of labels in covariate-shift," in *Proc. 31st Conference On Learning Theory*, 2018, pp. 1882–1886.

[55] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica, May*, vol. 23, p. 2016, 2016.

[56] D. B. Hunt, "Redlining," *Encyclopedia of Chicago*, 2005.

[57] N. M. Kinyanjui, T. Odonga, C. Cintas, N. C. Codella, R. Panda, P. Sattigeri, and K. R. Varshney, "Estimating skin tone and effects on classification performance in dermatology datasets," *arXiv preprint arXiv:1910.13268*, 2019.

[58] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[59] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Symp. on Security and Privacy*, 2016, pp. 598–617.

[60] 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.

[61] I. Chen, F. D. Johansson, and D. Sontag, "Why is my classifier discriminatory?" in *Advances in Neural Information Processing Systems*, 2018, pp. 3539–3550.

[62] H. Jiang, B. Kim, M. Guan, and M. Gupta, "To trust or not to trust a classifier," in *Advances in Neural Information Processing Systems*, 2018, pp. 5541–5552.

[63] N. Kallus and A. Zhou, "Residual unfairness in fair machine learning from prejudiced data," in *Proc. 35th International Conference on Machine Learning*, 2018, pp. 2439–2448.

[64] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.

[65] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proc. Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[66] A. Cotter, M. Gupta, and H. Narasimhan, "On making stochastic classifiers deterministic," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 910–10 920.

[67] 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.

[68] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, p. 0049124118782533, 2018.

[69] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proc. Conference on Fairness, Accountability, and Transparency*, 2019, p. 339–348.

[70] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[71] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[72] 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.

[73] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *Proc. 35th International Conference on Machine Learning*, 2018, pp. 60–69.

[74] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *Proc. of the 35th International Conference on Machine Learning*, 2018, pp. 2564–2572.

[75] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *Proc. 2018 IEEE Int. Symp. on Inf. Theory*, 2018, pp. 176–180.

[76] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness without demographics in repeated loss minimization," in *Proc. 35th International Conference on Machine Learning*, 2018, pp. 1929–1938.

[77] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 247–254.

[78] L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi, "Classification with fairness constraints: A meta-algorithm with provable guarantees," in *Proc. Conference on Fairness, Accountability, and Transparency*. ACM, 2019, pp. 319–328.

[79] W. Alghamdi, S. Asoodeh, H. Wang, F. P. Calmon, D. Wei, and K. N. Ramamurthy, "Model projection: Theory and applications to fair machine learning," in *Proc. 2020 IEEE Int. Symp. on Inf. Theory*, 2020.

[80] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, 2017, pp. 4066–4076.

[81] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[82] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proc. Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 1931–1940.

[83] L. Hu and Y. Chen, "Welfare and distributional impacts of fair classification," *arXiv preprint arXiv:1807.01134*, 2018.

[84] S. Chiappa, "Path-specific counterfactual fairness," in *Proc. Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7801–7808.

[85] Z. Lipton, J. McAuley, and A. Chouldechova, "Does mitigating ml's impact disparity require treatment disparity?" in *Advances in Neural Information Processing Systems*, 2018, pp. 8125–8135.

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

[87] A. Blum and T. Lykouris, "Advancing subgroup fairness via sleeping experts," *arXiv preprint arXiv:1909.08375*, 2019.

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

[89] 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.

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

[91] I. Sason and S. Verdu, "f-divergence inequalities," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[92] J. Liu, P. Cuff, and S. Verdú, "E_γ-resolvability," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 2629–2658, 2016.

[93] M. H. DeGroot, "Uncertainty, information, and sequential experiments," *The Annals of Mathematical Statistics*, vol. 33, no. 2, pp. 404–419, 1962.

[94] 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.

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

[96] 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.

[97] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, 1993.

[98] M. Anthony and P. L. Bartlett, *Neural network learning: Theoretical foundations*. Cambridge University press, 2009.

[99] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[100] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[101] T. Kanamori, T. Suzuki, and M. Sugiyama, "f-divergence estimation and two-sample homogeneity test under semiparametric density-ratio models," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 708–720, 2011.

[102] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[103] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[104] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Trans. Inf. Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[105] Y. Polyanskiy, H. V. Poor, and S. Verdú, "Channel coding rate in the finite blocklength regime," *IEEE Trans. Inf. Theory*, vol. 56, no. 5, pp. 2307–2359, 2010.

[106] N. Sharma and N. A. Warsi, "Fundamental bound on the reliability of quantum information transmission," *Physical review letters*, vol. 110, no. 8, p. 080501, 2013.

[107] K. Marton, "A measure concentration inequality for contracting markov chains," *Geometric & Functional Analysis GAFA*, vol. 6, no. 3, pp. 556–571, 1996.

[108] M. Raginsky, "Strong data processing inequalities and Φ-Sobolev inequalities for discrete channels," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[109] C. D. Aliprantis and K. C. Border, *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2006.

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

[111] R. T. Rockafellar and R. J. Wets, "On the interchange of subdifferentiation and conditional expectation for convex functionals," *Stochastics: An International Journal of Probability and Stochastic Processes*, vol. 7, no. 3, pp. 173–182, 1982.
[112] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.
[113] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[114] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, and R. D. Hjelm, "Mine: mutual information neural estimation," *arXiv preprint arXiv:1801.04062*, 2018.
[115] H. Hsu, S. Asoodeh, and F. P. Calmon, "Obfuscation via information density estimation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 906–917.

Hao Wang is currently a Ph.D. candidate at Harvard University. He received his M.S. degree in applied mathematics from Harvard University in 2019. In 2016, he received his B.S. degree in mathematics from the University of Science and Technology of China (USTC). He was awarded the 35th Guo Moruo Scholarship which is the highest honor at USTC. His research interests include information theory, statistical learning theory, and trustworthy machine learning.

Hsiang Hsu is currently a Ph.D. candidate in the Department of Computer Science at Harvard University, and a Facebook Fellow. He received B.S. degrees in Electrical Engineering and Mathematics, and the M.S. degree in Communication Engineering from National Taiwan University (NTU), Taipei, Taiwan, in 2014 and 2016, respectively. His research interests are in information theory and statistics, with applications to privacy, fairness, representation learning, and continual learning in machine learning.

Mario Diaz (M'21) was born in Guadalajara, Mexico, in 1988. He received the B.Eng. degree in electrical engineering from Universidad de Guadalajara, Guadalajara, Mexico, in 2011, the M.Sc. degree in probability and statistics from Centro de Investigación en Matemáticas, Guanajuato, Mexico, in 2013, and the Ph.D. degree in mathematics and statistics from Queen's University, Kingston, Canada, in 2017. He is currently a Research Associate in the Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas (IIMAS) at Universidad Nacional Autónoma de México, Mexico City, Mexico. Prior to this, he was a Postdoctoral Scholar at Arizona State University, Centro de Investigación en Matemáticas and Harvard University. His research interests include the mathematical and statistical foundations of information privacy, theoretical machine learning and random matrix theory.

Flavio P. Calmon is an Assistant Professor of Electrical Engineering at Harvard's John A. Paulson School of Engineering and Applied Sciences. Before joining Harvard, he was the inaugural data science for social good post-doctoral fellow at IBM Research in Yorktown Heights, New York. He received his Ph.D. in Electrical Engineering and Computer Science at MIT. His main research interests are information theory, inference, and statistics, with applications to fairness, privacy, machine learning, and communications engineering. Prof. Calmon has received the NSF CAREER Award, the Google Research Faculty Award, the Amazon Research Award, the IBM Open Collaborative Research Award, and Harvard's Lemann Brazil Research Fund Award.