Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis

Flavio du Pin Calmon, *Member, IEEE*, Dennis Wei, *Member, IEEE*, Bhanukiran Vinzamuri, *Member, IEEE*, Karthikeyan Natesan Ramamurthy, *Member, IEEE*, and Kush R. Varshney, *Senior Member, IEEE*

***Abstract*—Non-discrimination is a recognized objective in algorithmic decision making. In this paper, we introduce a novel probabilistic formulation of data pre-processing for reducing discrimination. We propose a convex optimization for learning a data transformation with three goals: controlling group discrimination, limiting distortion in individual data samples, and preserving utility. Several theoretical properties are established, including conditions for convexity, a characterization of the impact of limited sample size on discrimination and utility guarantees, and a connection between discrimination and estimation. Two instances of the proposed optimization are applied to datasets, including one on real-world criminal recidivism. Results show that discrimination can be greatly reduced at a small cost in classification accuracy and with precise control of individual distortion.**

***Index Terms*—Machine learning, ethics, optimization.**

I. Introduction

THIS paper relates to the problem of discrimination in the sense of prejudicial treatment of individuals based on membership in a legally protected group such as a race or gender. Making decisions explicitly on the basis of such protected attributes is referred to as direct discrimination or *disparate treatment*. More pervasive nowadays is indirect discrimination, in which protected attributes are not used but reliance on variables correlated with them leads to significantly different outcomes for different groups. The latter phenomenon is termed *disparate impact*. Indirect discrimination may be intentional, as in the historical practice of “redlining” [1] in the U.S. in which home mortgages were denied in zip codes populated primarily by minorities. However, the principle of disparate impact applies regardless of actual intent.

Discrimination has become an increasingly recognized problem in supervised machine learning as algorithms play larger roles in making decisions with major consequences on human lives, in areas from consumer finance to criminal justice. While supervised learning algorithms may appear at first to be fair and devoid of inherent bias, they in fact inherit any bias or discrimination present in the data on which they are trained [2]. Furthermore, simply removing protected variables from the data is not enough since it does nothing to address indirect discrimination and may in fact conceal it. The need for more sophisticated methods has made discrimination discovery and prevention an important research area [3]. One of the goals of this paper is to bring the problem to the attention of the signal processing and information theory communities in the hope of inspiring innovative solutions.

Algorithmic discrimination prevention approaches can be categorized as modifying one or more of the following to reduce the bias in decisions made by supervised learning methods: (a) the training data, (b) the learning algorithm, and (c) the ensuing decisions themselves. These are respectively classified as pre-processing [4], in-processing [5]–[7] and post-processing [8]. In this paper, we focus on pre-processing since it is the most flexible in terms of the data science pipeline: it allows any learning algorithm of choice to be used and can be integrated with data release and publishing mechanisms. In our view, pre-processing is also particularly amenable to signal processing and information theoretic ways of thinking. Traditionally, the processing of signals (data) in their original domain has been very much in the purview of signal processing, often referred to as “filtering” in its general sense. Pre-processing for fairness is one of the latest incarnations of this paradigm.

Researchers have also studied several notions of discrimination and fairness. Disparate impact is addressed by the principles of *statistical parity* and *group fairness* [9], which seek similar outcomes for all groups. In contrast, *individual fairness* [10] mandates that similar individuals be treated similarly irrespective of group membership. For classifiers and other predictive models, equal error rates for different groups are a desirable property [8], as is calibration or lack of *predictive bias* in the predictions [11]. The tension between the last two notions is described in [12], [13]; [14] is in a similar vein. Corbett-Davies *et al.* [15] discuss the trade-offs in satisfying prevailing notions of algorithmic fairness from a public safety standpoint. Since the present work pertains to pre-processing and not modeling, balanced error rates and predictive bias are less relevant criteria. Instead we focus on achieving both group and individual fairness where the latter is realized through constraints on distortion.

Manuscript received December 8, 2017; revised April 17, 2018 and July 22, 2018; accepted July 31, 2018. Date of publication August 17, 2018; date of current version September 27, 2018. This paper was presented in part at the 31st Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, December 2017. The guest editor coordinating the review of this paper and approving it for publication was Dr. Miguel Rodrigues. *(Corresponding author: Flavio du Pin Calmon.)*

F. P. Calmon is with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: flavio@seas.harvard.edu).

D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney are with IBM Research AI, Yorktown Heights, NY 10598 USA (e-mail: dwei@us.ibm.com; bhanu.vinzamuri@ibm.com; knatesa@us.ibm.com; krvarshn@us.ibm.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2018.2865887

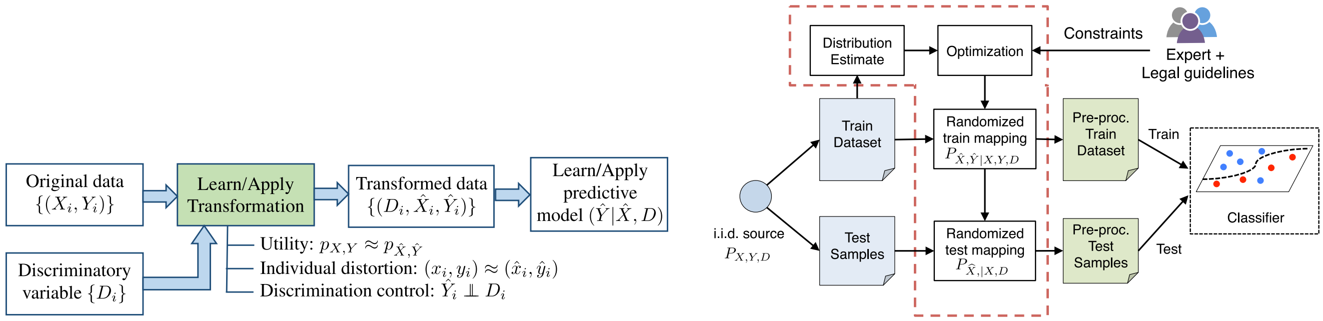
Original data $\{(X_i, Y_i)\}$
Learn/Apply Transformation
Transformed data $\{(D_i, \hat{X}_i, \hat{Y}_i)\}$
Learn/Apply predictive model $(\hat{Y}|\hat{X}, D)$
Discriminatory variable $\{D_i\}$
Utility: $p_{X,Y} \approx p_{\hat{X},\hat{Y}}$
Individual distortion: $(x_i, y_i) \approx (\hat{x}_i, \hat{y}_i)$
Discrimination control: $\hat{Y}_i \perp D_i$
Distribution Estimate
Optimization
Constraints
Expert + Legal guidelines
Train Dataset
Randomized train mapping $P_{\hat{X},\hat{Y}|X,Y,D}$
Pre-proc. Train Dataset
Train
Classifier
i.i.d. source $P_{X,Y,D}$
Test Samples
Randomized test mapping $P_{\hat{X},|X,D}$
Pre-proc. Test Samples
Test

Fig. 1. The proposed pipeline for supervised learning with discrimination prevention. *Learn* mode applies with training data and *apply* mode with novel test data. Note that test data also requires transformation before predictions can be obtained. The right-hand side figure provides a closer look at how the randomized mapping is used in practice.

Existing pre-processing approaches include sampling or reweighting the data to neutralize discriminatory effects [16], changing the individual data records [17], and using t-closeness [18] for discrimination control [19]. A common theme is the importance of balancing discrimination control against utility of the processed data. However, this prior work neither presents general and principled optimization frameworks for trading off these two criteria, nor allows connections to be made to the statistical signal processing and information theory literature via probabilistic descriptions. Another shortcoming is that individual distortion or fairness is not made explicit.

In this work, we (i) introduce a probabilistic framework for discrimination-preventing pre-processing in supervised learning, (ii) formulate an optimization problem for producing pre-processing transformations that trade off discrimination control, data utility, and individual distortion, (iii) characterize theoretical properties of the optimization approach (e.g., convexity, robustness to limited samples), and (iv) benchmark the ensuing pre-processing transformations on real-world datasets. Our aim in part is to work toward a more unified view of existing pre-processing concepts and methods, which may help to suggest refinements. While discrimination and utility are defined at the level of probability distributions, distortion is controlled on a per-sample basis, thereby limiting the effect of the transformation on individuals and ensuring a degree of individual fairness. Figure 1 illustrates the supervised learning pipeline that includes our proposed discrimination-preventing pre-processing.

The work of Zemel *et al.* [20] is closest to ours in also presenting a framework with three criteria related to discrimination control (group fairness), individual fairness, and utility. However, our formulation more naturally and generally encodes these desiderata. In [20], discrimination control is posed in terms of intermediate features rather than outcomes, making parameter selection less clear, and individual distortion does not take outcomes into account (being an ℓ_2-norm between original and transformed features). In addition, the proposal of [20] is a combination of pre-processing and in-processing since both the intermediate representation and the utility measure are specific to a particular cluster-based classifier. Lastly, [20] does not consider non-binary or multiple protected attributes.

The optimization approach is inspired by the information-theoretic privacy literature [18], [21]–[24] and, more broadly, rate-distortion theory [25], [26]—in fact, randomized pre-processing transformations for discrimination prevention share similarities with test channels in rate-distortion theory. The connection between privacy and discrimination control was noted in [10]. Whereas in privacy the output of an adversary's estimator is made (approximately) invariant to the private information *for any estimator chosen by the adversary*, in discrimination control the output of a classifier is made invariant to the protected variable. Since in the data pre-processing setting we do not assume any specific model used for classification, the invariance guarantee should hold for *all classifiers*, harking back to privacy. Other constraints in discrimination control (e.g. individual fairness) do not exactly translate to those found in privacy settings. Nevertheless, there are many information-theoretic metrics and techniques that can be shared by both areas.

We consider our approach "information-theoretic" in the sense that it requires at least partial knowledge of the distribution of the data, and focuses on characterizing randomized mappings (channels) directly in this setting instead of more algorithmic aspects. The information-theoretic route enables the creation of an optimization formulation that quantifies operational trade-offs between distortion and fairness. This optimization can be solved in practice to determine perturbation mappings for assuring fairness in supervised learning, with generalization guarantees provided by standard finite-sample analysis found in the information theory literature. One of the main strengths that information theory can provide to the area of fairness in learning is its ability to look at problems at a more conceptual and fundamental level, rather than focusing immediately (and perhaps too soon) on all the practical issues. Information theory can shed light on the fundamental trade-offs and limits involved in preventing discrimination in machine learning, whereas signal processing can inform the design of methods that approach these limits.

In Section II, we present our formulation for discrimination-preventing pre-processing. Given its novelty, we devote more effort than usual to discussing its motivations and potential variations. In Section III, we discuss several theoretical properties, including conditions under which the proposed optimization problem is convex, the generalization of discrimination and distortion guarantees from training to test data, and a connection between discrimination and estimation. We also characterize the possible degradation in discrimination and utility guarantees in terms of the training sample size under the common scenario where the true data distribution is estimated empirically from the training sample.

In Section IV, we demonstrate our framework by applying specific instances of it to a prison recidivism dataset [27] and the UCI Adult dataset [28]. This shows that discrimination, distortion, and utility loss can be controlled simultaneously with real data. Of note, the proposed pre-processing method is observed to reduce discrimination when training standard classifiers, particularly when compared to using the original data with and without removing protected variables. The decrease in discrimination is achieved with only a small loss in accuracy and strict bounds on individual distortion. We also show examples of pre-processing transformations, their effects on the datasets, and the demographic patterns that they reveal. Section V concludes the paper.

II. General Formulation

We are given a dataset consisting of n i.i.d. samples $\{(D_i, X_i, Y_i)\}_{i=1}^n$ from a joint distribution $p_{D,X,Y}$ with domain $\mathcal{D} \times \mathcal{X} \times \mathcal{Y}$. Here D denotes one or more protected (discriminatory) variables such as gender and race, X denotes other non-protected variables used for decision making, and Y is an *outcome* random variable. We use the term 'discriminatory' interchangeably with 'protected,' and not in the usual statistical sense. For instance, Y_i could represent a loan approval decision for individual i based on demographic information D_i and credit score X_i. We focus in this paper on discrete (or discretized) and finite domains $\mathcal{D}$ and $\mathcal{X}$ and binary outcomes, i.e. $\mathcal{Y} = \{0, 1\}$. There is no restriction on the dimensions of D and X.

Our goal is to determine a randomized mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ that (i) transforms the given dataset into a new dataset $\{(D_i, \hat{X}_i, \hat{Y}_i)\}_{i=1}^n$ which may be used to train a model, and (ii) similarly transforms data to which the model is applied, i.e., test data. Each $(\hat{X}_i, \hat{Y}_i)$ is drawn independently from the same domain $\mathcal{X} \times \mathcal{Y}$ as X, Y by applying $p_{\hat{X},\hat{Y}|X,Y,D}$ to the corresponding triplet (D_i, X_i, Y_i). Since D_i is retained as-is, we do not include it in the mapping to be determined. Motivation for retaining D is discussed later in Section III. For test samples, Y_i is not available at the input while $\hat{Y}_i$ may not be needed at the output. In this case, a reduced mapping $p_{\hat{X}|X,D}$ is used as given later in (7).

It is assumed that $p_{D,X,Y}$ is known along with its marginals and conditionals. This assumption is often satisfied using the empirical distribution of $\{(D_i, X_i, Y_i)\}_{i=1}^n$. In Section III, we state a result ensuring that discrimination and utility loss continue to be controlled if the distribution used to determine $p_{\hat{X},\hat{Y}|X,Y,D}$ differs from the distribution of test samples.

We propose that the mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ satisfy the three following properties.

A. Discrimination Control

The first objective is to limit the dependence of the transformed outcome $\hat{Y}$ on the protected variables D. We propose two alternative formulations. The first requires the conditional distribution $p_{\hat{Y}|D}$ to be close to a target distribution p_{Y_T} for all values of D,

$$J\left(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \leq \epsilon_{y,d}, \ \forall\, d \in \mathcal{D}, y \in \{0,1\}, \qquad (1)$$

where $J(\cdot,\cdot)$ denotes some distance function. In the second formulation, we constrain the conditional probability $p_{\hat{Y}|D}$ to be similar for any two values of D:

$$J\left(p_{\hat{Y}|D}(y|d_1), p_{\hat{Y}|D}(y|d_2)\right) \leq \epsilon_{y,d_1,d_2}, \quad \forall\, d_1, d_2 \in \mathcal{D}, y \in \{0,1\}. \qquad (2)$$

Note that the number of such constraints is $O(|\mathcal{D}|^2)$ as opposed to $O(|\mathcal{D}|)$ constraints in (1). The choice of p_{Y_T} in (1), and J and ϵ in (1) and (2) should be informed by legal formulations such as the "80% rule" [29], consultations with domain experts and stakeholders, and other societal considerations.

For this work, we choose J to be the following probability ratio measure:

$$J(p,q) = \left|\frac{p}{q} - 1\right|. \qquad (3)$$

This metric is motivated by the EEOC "80% rule" [29]. For example, $J(p_{Y|D}(1|0), p_{Y|D}(1|1)) \leq 0.2$ indicates that the fraction of outcomes $Y = 1$ for group $D = 0$ is within 80% of group $D = 1$. The combination of (3) and (1) generalizes the extended lift criterion proposed in the literature [30], while the combination of (3) and (2) generalizes selective and contrastive lift. The latter combination (2), (3) is used in the numerical results in Section IV. We note that the selection of a 'fair' target distribution p_{Y_T} in (1) is not straightforward; see Žliobaitė *et al.* [31] for one such proposal. Despite its practical motivation, we alert the reader that (3) may be unnecessarily restrictive on p when q is low.

In (1) and (2), discrimination control is imposed jointly with respect to all protected variables, e.g. all combinations of gender and race if D consists of those two variables. An alternative is to take the protected variables one at a time, and impose univariate discrimination control. In this work, we opt for the more stringent joint discrimination control, although legal formulations tend to be of the univariate type.

Formulations (1) and (2) control discrimination at the level of the overall population in the dataset. It is also possible to control discrimination within segments of the population by conditioning on additional variables B, where B is a subset of X. Constraint (1) would then generalize to

$$J\left(p_{\hat{Y}|D,B}(y|d,b), p_{Y_T|B}(y|b)\right) \leq \epsilon_{y,d,b}, \quad \forall\, d \in \mathcal{D}, y \in \{0,1\}, b \in \mathcal{B}.$$

Similar conditioning or 'context' for discrimination has been explored before in [17] in the setting of association rule mining. For example, B could represent the fraction of a pool of applicants that applied to a certain department, which enables the metric to avoid statistical traps such as Simpson's paradox [32]. One may wish to control for such variables in determining the presence of discrimination, while ensuring that population segments created by conditioning are large enough to derive statistically valid inferences. Moreover, we note that there may exist inaccessible latent variables that drive discrimination, and the metrics used here are inherently limited by the available data. Recent definitions of fairness that seek to mitigate this issue include [33]–[35]. We defer further investigation of causality and conditional discrimination to future work.

B. Distortion Control

The mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ should satisfy distortion constraints with respect to the domain $\mathcal{X} \times \mathcal{Y}$. These constraints restrict the mapping to reduce or avoid altogether certain large changes (e.g. a very low credit score being mapped to a very high credit score). Given a distortion metric $\delta : (\mathcal{X} \times \mathcal{Y})^2 \to \mathbb{R}_+$, we constrain the conditional expectation of the distortion as,

$$\mathbb{E}\left[\delta\big((x,y),(\hat{X},\hat{Y})\big) \mid D=d, X=x, Y=y\right] \leq c_{d,x,y} \quad \forall\,(d,x,y) \in \mathcal{D}\times\mathcal{X}\times\mathcal{Y}. \tag{4}$$

We assume that $\delta(x,y,x,y)=0$ for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$. Constraint (4) is formulated with pointwise conditioning on $(D,X,Y)=(d,x,y)$ in order to promote *individual* fairness. It ensures that distortion is controlled for every combination of (d,x,y), i.e. every individual in the original dataset, and more importantly, every individual to which a model is later applied. By way of contrast, an average-case measure in which an expectation is also taken over D,X,Y may result in high distortion for certain (d,x,y), likely those with low probability. Equation (4) also allows the level of control $c_{d,x,y}$ to depend on (d,x,y) if desired. We also note that (4) is a property only of the mapping $p_{\hat{X},\hat{Y}|D,X,Y}$, and does not depend on the distribution $p_{D,X,Y}$.

The expectation over $\hat{X},\hat{Y}$ in (4) encompasses several cases depending on the choices of the metric δ and thresholds $c_{d,x,y}$. If $c_{d,x,y}=0$, then no mappings with nonzero distortion are allowed for individuals with original values (d,x,y). If $c_{d,x,y}>0$, then certain mappings may still be disallowed by assigning them infinite distortion. Mappings with finite distortion are permissible subject to the budget $c_{d,x,y}$. Lastly, if δ is binary-valued (perhaps achieved by thresholding a multi-valued distortion function), it can be seen as classifying mappings into desirable ($\delta=0$) and undesirable ones ($\delta=1$). Here, (4) reduces to a bound on the conditional probability of an undesirable mapping, i.e.,

$$\Pr\left(\delta\big((x,y),(\hat{X},\hat{Y})\big)=1 \mid D=d, X=x, Y=y\right) \leq c_{d,x,y}. \tag{5}$$

C. Utility Preservation

In addition to constraints on individual distortions, we also require that the *distribution* of $(\hat{X},\hat{Y})$ be statistically close to the distribution of (X,Y). This is to ensure that a model learned from the transformed dataset (when averaged over the protected variables D) is not too different from one learned from the original dataset, e.g. a bank's existing policy for approving loans. For a given dissimilarity measure Δ between probability distributions (e.g. KL-divergence), we require that $\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$ be small. Distortion control and utility preservation, in the sense used in this paper, are intertwined: if $(\hat{X},\hat{Y})=(X,Y)$, then both perfect utility and zero distortion are achieved. We adopt the term "utility" to indicate the constraint $p_{\hat{X},\hat{Y}} \approx p_{X,Y}$, ensuring that a classifier learned on the transformed data will be close to one learned from the original distribution, hence "useful". Whereas distortion measures the similarity between individual data points, utility captures the preservation of the overall distribution.

D. Optimization Problem

Putting together the considerations from the three previous subsections, we arrive at the optimization problem below for determining a randomized transformation $p_{\hat{X},\hat{Y}|X,Y,D}$ mapping each sample (D_i,X_i,Y_i) to $(\hat{X}_i,\hat{Y}_i)$:

$$\begin{aligned}
\min_{p_{\hat{X},\hat{Y}|X,Y,D}} \quad & \Delta\left(p_{\hat{X},\hat{Y}}, p_{X,Y}\right) \\
\text{s.t.} \quad & J\left(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \leq \epsilon_{y,d}\ \forall\,(d,y)\in\mathcal{D}\times\mathcal{Y}, \\
& \mathbb{E}\left[\delta\big((x,y),(\hat{X},\hat{Y})\big) \mid D=d, X=x, Y=y\right] \\
& \quad \leq c_{d,x,y}\ \forall\,(d,x,y)\in\mathcal{D}\times\mathcal{X}\times\mathcal{Y}, \\
& p_{\hat{X},\hat{Y}|X,Y,D} \text{ is a valid distribution.}
\end{aligned} \tag{6}$$

We choose to minimize the utility loss Δ subject to constraints on individual distortion (4) and discrimination (we use (1) for concreteness, but (2) can be used instead), since it is more natural to place bounds on the latter two.

The distortion constraints (4) are an essential component of the problem formulation (6). Without (4) and assuming that $p_{Y_T}=p_Y$, it is possible to achieve perfect utility and non-discrimination simply by sampling $(\hat{X}_i,\hat{Y}_i)$ from the original distribution $p_{X,Y}$ independently of any inputs, i.e. $p_{\hat{X},\hat{Y}|X,Y,D}(\hat{x},\hat{y}|x,y,d)=p_{\hat{X},\hat{Y}}(\hat{x},\hat{y})=p_{X,Y}(\hat{x},\hat{y})$. Then $\Delta(p_{\hat{X},\hat{Y}},p_{X,Y})=0$, and $p_{\hat{Y}|D}(y|d)=p_{\hat{Y}}(y)=p_Y(y)=p_{Y_T}(y)$ for all $d\in\mathcal{D}$. Clearly, this solution is objectionable from the viewpoint of individual fairness, especially for individuals to whom a subsequent model is applied since it amounts to discarding an individual's data and replacing it with a random sample from the population $p_{X,Y}$. Constraint (4) seeks to prevent such gross deviations from occurring. The distortion constraints may conflict however with the discrimination constraint, in some cases rendering the optimization infeasible as illustrated in Section IV-C.

Conversely, a small distortion does not guarantee (in general) that the distribution is preserved (i.e., high statistical utility). For example, assuming $\mathcal{X},\mathcal{Y}\subseteq\mathbb{R}^m$ a reasonable distortion function is one that penalizes large deviations:

$$\delta\left((x,y),(\hat{x},\hat{y})\right)=\begin{cases}0, & \|x-\hat{x}\|\vee\|y-\hat{y}\|\leq\varepsilon,\\ C, & \text{otherwise,}\end{cases}$$

where C is a large constant. Here, zero distortion can be achieved by constraining $\hat{X},\hat{Y}$ to take values in an ε-net. If X,Y are continuous random variables (i.e. have a density) and Δ is

KL-divergence, then the distance between $p_{X,Y}$ and $p_{\hat{X},\hat{Y}}$ cannot be upper bounded.

Despite being convex, the number of constraints and variables in (6) scales with the cardinality of the alphabets $\mathcal{D}$, $\mathcal{X}$ and $\mathcal{Y}$. The exact complexity of the optimization depends on the solver (for an overview, see [36]). If the number of variables and constraints is large, a dimensionality reduction step may be necessary for the computational feasibility. Proposition 2 addresses this scalability in terms of generalization of the fairness/utility guarantees.

III. Theoretical Properties

We provide next a sequence of theoretical results regarding the optimization formulation (6). More specifically, we outline conditions under which the formulation is (quasi)convex, discuss the generalization of discrimination and distortion guarantees from training to test data, and prove a robustness result for pre-processing mappings obtained by solving (6) using the empirical distribution of the data. Finally, we discuss the theoretical connection between estimation and discrimination.

The convexity property in Proposition 1 ensures that the transformation returned by the optimization formulation does in fact achieve the highest utility subject to the distortion and fairness requirements. Convexity also enables (6) to be solved using standard convex solvers, as illustrated in the numerical experiments in Section IV. Proposition 2 quantifies how the train-time guarantees for fairness and utility generalize to unseen samples. The generalization error will depend on the number of samples used to estimate the distribution $p_{D,X,Y}$ and the cardinality of the alphabets.

A. Convexity

The next proposition provides conditions under which (6) is a convex or quasiconvex optimization problem and can thus be tractably solved to optimality. The proof is presented in Appendix A.

Proposition 1: Problem (6) is a (quasi)convex optimization if $\Delta(\cdot,\cdot)$ is (quasi)convex and $J(\cdot,\cdot)$ is quasiconvex in their respective first arguments (with the second arguments fixed). If discrimination constraint (2) is used in place of (1), then the condition on J is that it be jointly quasiconvex in both arguments.

B. Generalization of Guarantees to Test Data

The proposed pre-processing method has two modes of operation (Fig. 1): train and test. In train mode, the optimization problem (6) is solved with the training dataset as input to determine a mapping $p_{\hat{X},\hat{Y}|X,Y,D}$, which is then applied to the same training data. The resulting pre-processed data thus satisfies discrimination constraints (1) or (2) and distortion constraint (4). In test mode, new data points (X, D) are received (Y is not available) and transformed into $(\hat{X}, D)$ through a randomized mapping $p_{\hat{X}|X,D}$ given by marginalizing $p_{\hat{X},\hat{Y}|X,Y,D}$ over $Y, \hat{Y}$:

$$p_{\hat{X}|D,X}(\hat{x}|d,x) = \sum_{y,\hat{y}} p_{\hat{X},\hat{Y}|X,Y,D}(\hat{x},\hat{y}|x,y,d)p_{Y|X,D}(y|x,d). \tag{7}$$

This subsection discusses the generalization of train-time guarantees (1), (2), (4) to test data.

1) Distortion Control: The distortion constraint (4) changes as a consequence of the marginalization over Y in (7). More specifically, the bound on the expected distortion for each sample becomes

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}\left[\delta\big((x,Y),(\hat{X},\hat{Y})\big) \mid D = d, X = x, Y\right]\right]\\ \leq \sum_{y\in\mathcal{Y}} p_{Y|X,D}(y|x,d)c_{x,y,d} \triangleq c_{x,d}\,.\end{aligned} \tag{8}$$

If the distortion control values $c_{x,y,d}$ are independent of y, then (8) and (4) are in fact the same.

2) Discrimination Control When Protected Variables are Used: Recall from Section II that the proposed transformation retains the protected variables D. We first consider the case where models trained on the transformed data to predict $\hat{Y}$ are allowed to depend on D, i.e., a classification model $f_\theta(\hat{X}, D)$ that approximates $p_{\hat{Y}|\hat{X},D}$ is fit to the pre-processed training set, where θ are the parameters of the model. While such models may qualify as disparate treatment, the intent and effect is to better mitigate disparate impact resulting from the model. In this respect our proposal shares the same spirit with ‘fair’ affirmative action in Dwork *et al.* [10] (fairer on account of distortion constraint (4)).

Under the above assumption, let $\widetilde{Y}$ be the output of a model based on D and $\hat{X}$. To remove the separate issue of model accuracy, suppose for simplicity that the (possibly randomized) model provides a good approximation to the conditional distribution of $\hat{Y}$, i.e., $p_{\widetilde{Y}|\hat{X},D}(\widetilde{y}|\hat{x},d) \approx p_{\hat{Y}|\hat{X},D}(\widetilde{y}|\hat{x},d)$. Then for individuals in a protected group $D = d$, the conditional distribution of $\widetilde{Y}$ is given by

$$p_{\widetilde{Y}|D}(\widetilde{y}|d) = \sum_{\hat{x}} p_{\widetilde{Y}|\hat{X},D}(\widetilde{y}|\hat{x},d)p_{\hat{X}|D}(\hat{x}|d) \tag{9}$$

$$\approx \sum_{\hat{x}} p_{\hat{Y}|\hat{X},D}(\widetilde{y}|\hat{x},d)p_{\hat{X}|D}(\hat{x}|d) \tag{10}$$

$$= p_{\hat{Y}|D}(\widetilde{y}|d). \tag{11}$$

Hence the model output $p_{\widetilde{Y}|D}$ can also be controlled by (1) or (2).

3) Discrimination Control When Protected Variables are Suppressed: Suppose now that the protected variables D must be suppressed from the model input, perhaps to comply with legal requirements regarding their non-use. (We still assume that $\hat{X}$ may depend on both X and D.) Then a predictive model can depend only on $\hat{X}$ and approximate $p_{\hat{Y}|\hat{X}}$, i.e., $p_{\widetilde{Y}|\hat{X},D}(\widetilde{y}|\hat{x},d) = p_{\widetilde{Y}|\hat{X}}(\widetilde{y}|\hat{x}) \approx p_{\hat{Y}|\hat{X}}(\widetilde{y}|\hat{x})$. In this case we have

$$p_{\widetilde{Y}|D}(\widetilde{y}|d) \approx \sum_{\hat{x}} p_{\hat{Y}|\hat{X}}(\widetilde{y}|\hat{x})p_{\hat{X}|D}(\hat{x}|d), \tag{12}$$

which in general is not equal to $p_{\hat{Y}|D}(\widetilde{y}|d)$ in (11).

The quantity on the right-hand side of (12) is less straightforward to control and we leave a full treatment of this case to future work. Below we outline two approaches based on the observation that (12) becomes equivalent to (11) if the Markov relationship $D \to \hat{X} \to \hat{Y}$ (i.e., $p_{\hat{Y}|\hat{X},D} = p_{\hat{Y}|\hat{X}}$) holds. Thus train-time discrimination guarantees still hold for test samples if the additional constraint $p_{\hat{X},\hat{Y}|D,X,Y} = p_{\hat{Y}|\hat{X}} p_{\hat{X}|D,X,Y}$ is satisfied. We refer to (6) with the additional constraint $p_{\hat{X},\hat{Y}|D,X,Y} = p_{\hat{Y}|\hat{X}} p_{\hat{X}|D,X,Y}$ as the *suppressed optimization formulation* (SOF). Alas, since the added constraint is nonconvex, the SOF is not a convex program but it is convex in $p_{\hat{X}|D,X,Y}$ for a fixed $p_{\hat{Y}|\hat{X}}$ and vice-versa (i.e., it is biconvex). We propose next two strategies for addressing the SOF.

1) The first approach is to restrict $p_{\hat{Y}|\hat{X}} = p_{Y|X}$ and solve (6) for $p_{\hat{X}|D,X,Y}$. If $\Delta(\cdot,\cdot)$ is an f-divergence, then

$$\begin{aligned}\Delta\left(p_{X,Y}, p_{\hat{X},\hat{Y}}\right) &= D_f\left(p_{X,Y} \| p_{\hat{X},\hat{Y}}\right)\\ &= \sum_{x,y} p_{\hat{X},\hat{Y}}(x,y) f\left(\frac{p_{X,Y}(x,y)}{p_{\hat{X},\hat{Y}}(x,y)}\right)\\ &\geq \sum_{x} p_{\hat{X}}(x) f\left(\sum_{y} p_{\hat{Y}|\hat{X}}(y|x)\frac{p_{X,Y}(x,y)}{p_{\hat{X},\hat{Y}}(x,y)}\right)\\ &= D_f\left(p_X \| p_{\hat{X}}\right),\end{aligned}$$

where the inequality follows from convexity of f. Since the last quantity is achieved by setting $p_{\hat{Y}|\hat{X}} = p_{Y|X}$, this choice is optimal in terms of the objective function. It may, however, render the constraints in (6) infeasible. Assuming feasibility is maintained, this approach has the added benefit that a classifier $f_\theta(x) \approx p_{Y|X}(\cdot|x)$ can be trained using the original (non-perturbed) data, and maintained for classification at test time.

2) Alternatively, a solution can be found through alternating minimization: fix $p_{\hat{Y}|\hat{X}}$ and solve the SOF for $p_{\hat{X}|D,X,Y}$, and then fix $p_{\hat{X}|D,X,Y}$ at the optimal solution and solve the SOF for $p_{\hat{Y}|\hat{X}}$. The resulting sequence of values of the objective function is non-increasing, but may converge to a local minima.

C. Robustness to Mismatched Prior Distribution

Next we consider the case where the distribution $p_{D,X,Y}$ used to determine the transformation differs from the true distribution $q_{D,X,Y}$ of training and test samples. This occurs in particular when $p_{D,X,Y}$ is the empirical distribution computed from n i.i.d. samples of $q_{D,X,Y}$, which is not known exactly. In this situation, discrimination control and utility are still guaranteed for samples drawn from $q_{D,X,Y}$ that are transformed using $p_{\hat{Y},\hat{X}|X,Y,D}$, where the latter is obtained by solving (6) with $p_{D,X,Y}$. Note that (6) ensures the distortion control constraint (4) is satisfied regardless of data distribution. Denoting by $q_{\hat{Y}|D}$ and $q_{\hat{X},\hat{Y}}$ the corresponding distributions for $\hat{Y}, \hat{X}$ and D when $q_{D,X,Y}$ is transformed using $p_{\hat{Y},\hat{X}|X,Y,D}$, we have $J(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)) \rightarrow J(q_{\hat{Y}|D}(y|d), p_{Y_T}(y))$ and $\Delta(p_{X,Y}, p_{\hat{X},\hat{Y}}) \rightarrow \Delta(q_{X,Y}, q_{\hat{X},\hat{Y}})$ for n sufficiently large. The next proposition provides an estimate of the rate of this convergence in terms of n and assuming $p_{Y,D}(y,d)$ is fixed and bounded away from zero. Its proof can be found in Appendix B.

Proposition 2: Let $p_{D,X,Y}$ be the empirical distribution obtained from n i.i.d. samples that is used to determine the mapping $p_{\hat{Y},\hat{X}|X,Y,D}$, and $q_{D,X,Y}$ be the true distribution of the data, with support size $m \triangleq |\mathcal{X} \times \mathcal{Y} \times \mathcal{D}|$. In addition, denote by $q_{D,\hat{X},\hat{Y}}$ the joint distribution after applying $p_{\hat{Y},\hat{X}|X,Y,D}$ to samples from $q_{D,X,Y}$. If for all $y \in \mathcal{Y}$, $d \in \mathcal{D}$ we have $p_{Y,D}(y,d) > 0$, $p^* \triangleq \min_{y\in\mathcal{Y},d\in\mathcal{D}} p_{Y_T}(y) p_D(d)$, $J(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)) \leq \epsilon$, where J is given in (3), and

$$\Delta\left(p_{X,Y}, p_{\hat{X},\hat{Y}}\right) = \sum_{x,y}\left|p_{X,Y}(x,y) - p_{\hat{X},\hat{Y}}(x,y)\right| \leq \mu,$$

then with probability $1-2\beta$,

$$J\left(q_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) - \epsilon \leq \frac{1}{p^*}\sqrt{\frac{8}{n}\left(\ln\frac{1}{\beta} + m\right)} \quad (13)$$

$$\Delta\left(q_{X,Y}, q_{\hat{X},\hat{Y}}\right) - \mu \leq \sqrt{\frac{8}{n}\left(\ln\frac{1}{\beta} + m\right)}. \quad (14)$$

Proposition 2 guarantees that, as long as n is sufficiently large, the utility and discrimination control guarantees will approximately hold when $p_{\hat{X},\hat{Y}|Y,X,D}$ is applied to fresh samples drawn from $q_{D,X,Y}$. In particular, the utility and discrimination guarantees will converge to the ones used as parameters in the optimization at a rate of at least $O(\sqrt{\frac{1}{n}})$. The convergence rate for the utility guarantee is tied to the support size, and for large m a dimensionality reduction step may be required to better control the convergence. A bound with the same asymptotic behavior holds for discrimination constraints of the form (2).

D. On Estimation and Discrimination

There is a close relationship between estimation and discrimination. If the protected variable D can be reliably estimated from the outcome variable Y, then it is reasonable to expect that the discrimination control constraint (1) does not hold for small values of $\epsilon_{y,d}$. We make this intuition precise in the case when J is given in (3) next.

More specifically, we prove that if the advantage of estimating D from Y over a random guess is large, then there must exist a value of d and y such that $J(p_{Y|D}(y|d), p_{Y_T}(y))$ is also large. Thus, standard estimation methods can be used to detect the presence of discrimination: if an estimation algorithm can estimate D from Y, then discrimination may be present. Alternatively, if discrimination control is successful, then no estimator can significantly improve upon a random guess when estimating D from Y.

We denote the highest probability of correctly guessing D from an observation of Y by $P_c(D|Y)$, where

$$P_c(D|Y) \triangleq \max_{D\rightarrow Y\rightarrow \hat{D}} \Pr\left(D = \hat{D}\right), \quad (15)$$

and the maximum is taken across all estimators $p_{\hat{D}|Y}$ that satisfy the Markov condition $D \rightarrow Y \rightarrow \hat{D}$. For D and Y defined over finite supports, this is achieved by the maximum *a posteriori* (MAP) estimator and, consequently,

$$P_c(D|Y) = \sum_{y\in\mathcal{Y}} p_Y(y) \max_{d\in\mathcal{D}} p_{D|Y}(d|y). \quad (16)$$

Let p_D^* be the most likely outcome of D, i.e., $p_D^* \triangleq \max_{d \in \mathcal{D}} p_D(d)$. The (multiplicative) advantage over a random guess is given by

$$\mathsf{Adv}(D|Y) \triangleq \frac{P_c(D|Y)}{p_D^*}. \tag{17}$$

The next proposition connects discrimination and estimation. Simply put, it shows that if a protected variable D can be reliably estimated from the decision variable Y, then Y can discriminate in terms of D. The proposition is given in terms of discrimination measured as (1), but a similar result holds for (2). The proof is given in Appendix C.

Proposition 3: For D and Y defined over finite support sets, if $\mathsf{Adv}(D|Y) > 1 + \epsilon$, then for any p_{Y_T}, there exists $y \in \mathcal{Y}$ and $d \in \mathcal{D}$ such that $|\frac{p_{Y|D}(y|d)}{p_{Y_T}(y)} - 1| > \epsilon$.

Remark 1: The previous proposition demonstrates that if there is a strong correlation between D and Y, i.e., $P_c(D|Y)$ is large and the protected variable D can be easily estimated from Y, fairness cannot be achieved in terms of the metric (3). In this case, there is an unfavorable trade-off between the fairness and distortion objective, the nature of which depends on the chosen distortion function.

IV. Experimental Results

This section illustrates the application of the data pre-processing pipeline in Fig. 1 to two real-world datasets, ProPublica's COMPAS recidivism data [27] and the UCI Adult dataset [28]. Section IV-A provides details on the datasets while Section IV-B describes how the general formulation in Section II is instantiated for each dataset. We present the trade-offs obtained by our optimization approach, first between discrimination and utility in Section IV-C, and then between discrimination and classification accuracy in Section IV-D when the pre-processed data is used to train standard prediction models. We also discuss in Sections IV-E and IV-F the pre-processing transformations produced by our formulation, their effects on the datasets, and the patterns of societal bias that they capture.

A. Data

The recidivism data [27] that we use was published by ProPublica as part of their investigation [37] into racial bias exhibited by Northpointe's COMPAS algorithm [38], a proprietary tool used in some US jurisdictions to score incarcerated individuals on their risk of reoffending. The investigation touched off a well-publicized debate around the COMPAS algorithm; see e.g. [13] for a technical analysis of how unequal error rates between African-Americans and Caucasian-Americans are a consequence of the a priori higher prevalence of recidivism among African-Americans and the calibration of the model. In this work, our interest is not in the COMPAS algorithm but rather in the underlying recidivism records. Using the proposed pre-processing approach, we demonstrate the technical feasibility of mitigating the disparate impact of these rearrests on different demographic groups while also preserving utility and individual fairness. (We do not address the associated societal implications.)

TABLE I
Recidivism Dataset Features

Feature	Values	Comments
2-year recidivism	$\{0, 1\}$	1 if re-offended, 0 otherwise
Gender	{Male, Female}	
Race	{Caucasian, African-American}	Races with small samples removed
Age category	$\{< 25, 25-45, > 45\}$	Years of age
Charge degree	{Felony, Misdemeanor}	For the current arrest
Prior counts	$\{0, 1-3, > 3\}$	Number of prior crimes

The outcome variable Y in the recidivism data is a binary indicator of whether an individual re-offended within two years of release. Thus we filtered data from Broward County to include only people who had either recidivated within two years or had at least two years outside of a correctional facility. Instances with missing data were also removed, leaving 5278 instances for our analysis. Both race and gender were considered as protected variables (D), and other features selected were severity of charge, number of prior crimes, and age category (X). Table I shows the encoding of these variables.

For the UCI Adult dataset (32561 instances), the features were categorized as protected variables (D): gender (male, female) and race (white, minority); decision variables (X): age (quantized to decades) and education (quantized to years); and response variable (Y): income (binary, high or low). While the response variable considered here is income, the dataset could be regarded as a simplified proxy for analyzing other financial outcomes such as credit approvals.

B. Specific Instantiations of Formulation

In all experiments, we approximate $p_{D,X,Y}$ using the empirical distribution of (D, X, Y) in the data and solve (6) using a standard convex solver [39]. We then apply the optimized randomized mapping $p_{\hat{X},\hat{Y}|D,X,Y}$ independently to each data sample to obtain a pre-processed dataset.

For utility metrics Δ, we use both the total variation distance, i.e. $\Delta(p_{X,Y}, p_{\hat{X},\hat{Y}}) = \frac{1}{2}\sum_{x,y} |p_{X,Y}(x,y) - p_{\hat{X},\hat{Y}}(x,y)|$, as well as KL divergence $D_{\mathsf{KL}}(p_{X,Y} \| p_{\hat{X},\hat{Y}})$, in part to demonstrate the versatility of our formulation. For the discrimination constraint, we use the combination of (2) and (3) (except in Section IV-F where (1) is used in place of (2)) with a single parameter on the right-hand side, $\epsilon_{y,d_1,d_2} = \epsilon$. The distortion function δ is chosen differently for the two datasets as described below, based on the differing semantics of the variables in the two applications. The specific numerical values were chosen for demonstration purposes to be reasonable to our judgment and can easily be changed according to the requirements of a domain user. We emphasize that the distortion values were not selected to optimize the trade-offs in Sections IV-C and IV-D.

Distortion function for Recidivism: We use the expected distortion constraint in (4) with $c_{d,x,y}$ as specified later. The distortion function δ has the following behavior. Jumps of more than one category in age and prior counts are heavily discouraged by a high distortion penalty (10^4) for such transformations. We impose the same penalty on increases in recidivism (change of Y from 0 to 1). Both these choices are made in the interest of

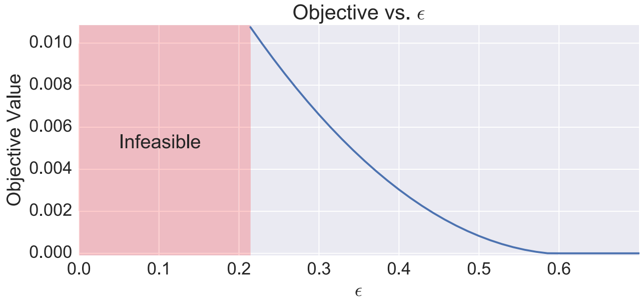

Fig. 2. KL divergence vs. discrimination parameter ϵ for the recividism dataset and distortion parameter $c = 0.25$.

individual fairness. Furthermore, for every jump to an adjacent category for age and prior counts, a penalty of 1 is assessed, and a similar jump in charge degree incurs a penalty of 2. Reduction in recidivism (1 to 0) has a penalty of 2. The total distortion for each individual is the sum of squared distortions for each attribute of X.

Distortion function for Adult: We use three conditional probability constraints of the form in (5). In constraint i, the distortion function returns 1 in case (i) and 0 otherwise: (1) if income is decreased, age is not changed and education is increased by at most 1 year, (2) if age is changed by a decade and education is increased by at most 1 year regardless of the change of income, (3) if age is changed by more than a decade or education is lowered by any amount or increased by more than 1 year. The corresponding probability bounds $c_{d,x,y}$ are $0.1, 0.05, 0$ (no dependence on d, x, y). As a consequence, and in the same broad spirit as in the recidivism application, decreases in income, small changes in age, and small increases in education (events (1), (2)) are permitted with small probabilities, while larger changes in age and education (event (3)) are not allowed at all.

C. Discrimination-Utility Trade-Off

As a first illustration, we consider the recividism dataset with KL divergence as the utility metric, distortion parameter $c_{d,x,y} = c = 0.25$, and all other choices as described in Section IV-B. We computed the minimal KL divergence resulting from solving (6) for different values of the discrimination control parameter ϵ. Fig. 2 shows the resulting trade-off between utility and discrimination. Around $\epsilon = 0.2$, no feasible solution can be found that also satisfies the distortion constraint. Above $\epsilon = 0.59$, the discrimination control is loose enough to be satisfied by the original dataset with just an identity mapping ($D_{\mathsf{KL}}(p_{X,Y} \| p_{\hat{X},\hat{Y}}) = 0$). In between, the optimal value varies as a smooth function.

D. Discrimination-Accuracy Trade-Off Compared to Baseline Methods

Next we evaluate the accuracy of classifiers trained on pre-processed data satisfying different discrimination levels. Classification accuracy represents a step beyond the utility optimized in (6), which is a distance between data distributions and thus an indirect measure. For this purpose, the datasets are split into training and test sets via 5-fold cross-validation. The training sets are pre-processed according to Section IV-B, this time using total variation as the utility metric and two values for the discrimination parameter, $\epsilon = 0.05, 0.10$. For the recidivism dataset, the distortion parameter is set to $c_d = 0.4, 0.3$ for d corresponding to African-Americans and Caucasians respectively. Two classifiers are fit to the pre-processed data: logistic regression (LR) and random forest (RF). We chose LR and RF since they are standard classification algorithms used in data analysis, but other classifiers can also be used instead.

For the test set, we first compute the test-time mapping $p_{\hat{X}|D,X}$ in (7) using $p_{\hat{X},\hat{Y}|D,X,Y}$ and $p_{Y|X,D}$ estimated from the training set. We then independently transform each test sample (d_i, x_i) using $p_{\hat{X}|D,X}$, preserving the protected variable D, i.e. $(d_i, x_i) \xrightarrow{p_{\hat{X}|D,X}} (d_i, \hat{x}_i)$. Each trained classifier f is applied to the transformed test samples, obtaining outputs $\widetilde{y}_i = f(d_i, \hat{x}_i)$ which are evaluated against y_i.

Our proposed approach is benchmarked against two baselines, leaving the dataset as-is during training and testing, and suppressing the protected variable D, again during both training and testing. We also compare against the *learning fair representations* (LFR) algorithm from [20]. Due to the lack of available code, we implemented LFR ourselves in Python and solved the associated optimization problem using the SciPy package [40]. The parameters for LFR were set as recommended in [20]: $A_z = 50$ (group fairness), $A_x = 0.01$ (individual fairness), and $A_y = 1$ (prediction accuracy), the last after tuning over the set $\{0.1, 0.5, 1, 5, 10\}$ to maximize prediction accuracy. The results did not significantly change within a reasonable variation of these three parameters. The number of prototypes K was set to 10.

As discussed in the introduction, LFR has fundamental differences from the proposed framework. In particular, LFR only considers binary-valued D, and consequently, we restrict D to be binary in this subsection, specifically race for recidivism and gender for Adult, and drop the other protected variable (gender for recidivism and race for Adult). However, our method is *not* restricted to D being binary or univariate and we consider race and gender jointly in the other experiments.

We report the trade-off between two metrics: (i) the empirical discrimination of the classifier on the test set, given by $\max_{d,d' \in \mathcal{D}} J(p_{\widetilde{Y}|D}(1|d), p_{\widetilde{Y}|D}(1|d'))$, where $p_{\widetilde{Y}|D}(1|d) = \frac{1}{n_d} \sum_{\{\hat{x}_i, d_i\}: d_i = d} f(d_i, \hat{x}_i)$ is the empirical conditional distribution and n_d is the number of samples with $d_i = d$; (ii) the empirical accuracy, measured by the Area under ROC (AUC) of $\widetilde{y}_i = f(d_i, \hat{x}_i)$ compared to y_i, using 5-fold cross validation. Fig. 3 presents the operating points achieved by each procedure in the discrimination-accuracy space defined by these metrics. For the recidivism dataset, there is significant discrimination in the original data, which is reflected by both LR and RF when the data is not transformed. Dropping the D variable reduces discrimination with a negligible impact on classification. However, discrimination is far from removed since the features X are correlated with D, i.e., there is indirect discrimination. LFR with the recommended parameters is successful in further reducing discrimination while still achieving high prediction performance for the task.

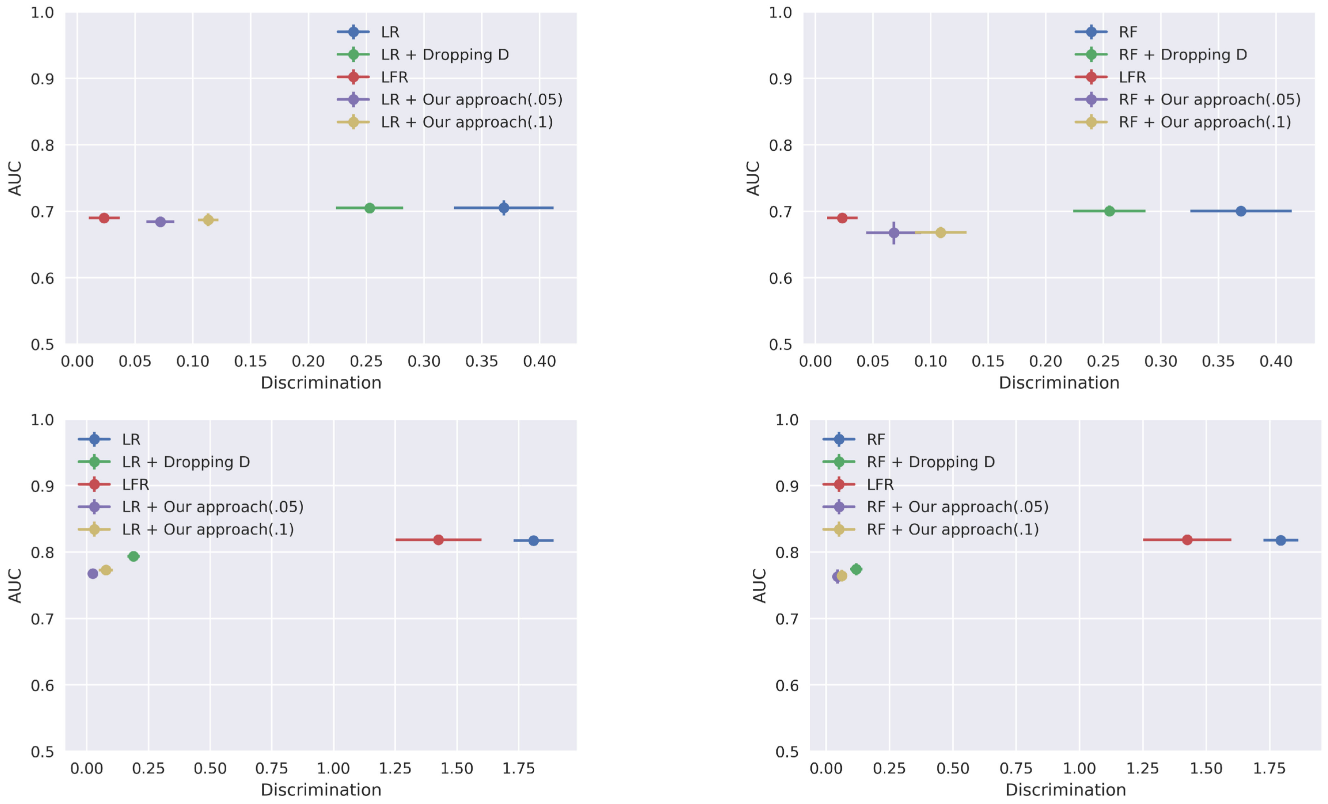

Fig. 3. Discrimination-AUC plots for two different classifiers. Top row is for recidivism dataset, and bottom row for UCI Adult dataset. First column is logistic regression (LR), and second column is random forests (RF).

Our proposed optimized pre-processing successfully decreases the empirical discrimination close to the target ϵ values of 0.05 and 0.10 (x-axis). Deviations are expected due to the approximation of $\hat{Y}$, the output of the transformation, by $\widetilde{Y}$, the output of each classifier, and also due to the randomized nature of the method. The decreased discrimination comes at an accuracy cost, which is greater in this case than for LFR. A possible explanation is that LFR is free to search across different representations whereas our method preserves the domain of the original variables and more importantly is restricted by the chosen distortion metric. In the recividism application, we heavily penalize increases in recidivism from 0 to 1 as well as large changes in prior counts and age. When combined with the other constraints in the optimization, this may alter the joint distribution after pre-processing and by extension the classifier output. Accuracy could be increased by relaxing the distortion constraint as long as this is acceptable to the domain user. We highlight again that the distortion metric was not chosen to optimize the trade-off in Fig. 3.

For the Adult dataset, dropping the protected variable does significantly reduce discrimination, in contrast with the recividism dataset. Our method further reduces discrimination towards the target ϵ values. The loss of prediction performance is again due to satisfying the distortion and discrimination constraints. On the other hand, LFR with the recommended parameters provides only a small reduction in discrimination. This does not contradict the results in [20] since here we have adopted a multiplicative discrimination metric (3) whereas [20] used an additive metric. Moreover, we reduced the Adult dataset to 31 binary features which is different from [20] where they additionally considered the test dataset for Adult (12661 instances) also and created 103 binary features. By varying the LFR parameters, it is possible to attain low empirical discrimination but with a large loss in prediction performance (below the plotted range).

In light of Fig. 3 and the differences in distortion and discrimination control, it cannot be said that either our method or LFR outperforms the other in trading off discrimination versus accuracy. In our approach however, discrimination can be controlled directly and transparently by setting the parameter ϵ, and changes to individual features are also finely controlled by the distortion metric. Both of these relationships are less clear with LFR.

E. Pre-Processing Transformation and Output for Recidivism Data

In the remainder of Section IV, we take a closer look at the pre-preocessing transformations produced by solving (6) and their effects on the datasets. For the results in this subsection on the recidivism data, we return to using KL divergence as the utility measure. The discrimination and distortion control parameters were set as $\epsilon = 0.1$ and $c = 0.5$. The corresponding optimal utility (KL divergence) was 0.021.

In general, the mappings $p_{\hat{X},\hat{Y}|X,Y,D}$ resulting from (6) can reveal insights on the nature of disparate impact and how to mitigate it. We illustrate this on the recidivism dataset. Fig. 4 displays the mapping restricted to certain socio-demographic

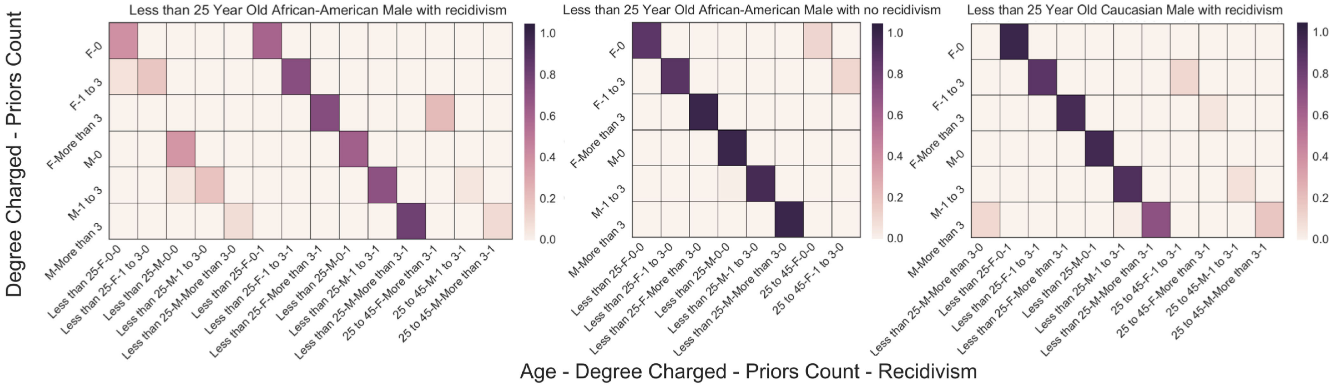

Fig. 4. Pre-processing mappings $p_{\hat{X},\hat{Y}|X,Y,D}$ from the recividism data with $\epsilon = 0.1$ and $c = 0.5$ for: (**left**) $D =$ (African-American, Male), less than 25 years (X), $Y = 1$, (**middle**) $D =$ (African-American, Male), less than 25 years (X), $Y = 0$, and (**right**) $D =$ (Caucasian, Male), less than 25 years (X), $Y = 1$. Original charge degree and prior counts (X) are shown in vertical axis, while the transformed age category, charge degree, prior counts and recidivism ($\hat{X}, \hat{Y}$) are represented along the horizontal axis. The charge degree F indicates felony and M indicates misdemeanor. Colors indicate mapping probability values. Columns included only if the sum of its values exceeds 0.05.

groups. First consider young African-American males (leftmost plot). This group has a high recidivism rate, and hence the most prominent action of the mapping (besides the identity transformation) is to change the recidivism value from 1 to 0. The frequency of this event however is lower for current charges that are felonies and for higher prior counts. The next most prominent action is to change the age category from young to middle-aged (25 to 45 years). This also effectively reduces the average value of $\hat{Y}$ for young African-American males by moving individuals out of the age category. Furthermore, the mapping for young African-American males who do not recidivate (middle plot) is essentially the identity mapping, with the exception of some age increases. This is expected since increasing recidivism is heavily penalized. For young Caucasian males who recidivate (right plot), the actions of the proposed transformation are similar to those for young African-American males who recidivate, i.e., either the outcome variable is changed to 0 or the age category is increased. However the probabilities of the transformations are lower since Caucasian males have, according to the dataset, a lower recidivism rate.

We applied the mapping shown partially in Fig. 4 to the dataset (a single realization of the randomization). First, to simply verify that discrimination control was achieved as expected, we examine the dependence of the outcome variable on the discrimination variable before and after the transformation. The corresponding conditionals $p_{Y|D}$ and $p_{\hat{Y}|D}$ are illustrated in Table II, where clearly $\hat{Y}$ is less dependent on D compared to Y. More precisely, the values of $p_{\hat{Y}|D}(1|d)$ are indeed controlled to within $\epsilon = 0.1$. Since, as mentioned, increases in recidivism are heavily penalized, the net effect of the transformation is to decrease the recidivism risk of males, and particularly African-American males.

A more detailed view of the pre-processed data is shown in Fig. 5, specifically the changes in recidivism rates (bottom panels) from the original rates (top panels) as a function of the features X and group D. For the overall population (leftmost column), the changes in recidivism rates are all negative, again a reflection of the distortion constraint that effectively disallows changing the outcome to 1. The maximum decreases are observed for African-American males since they have the highest value of $p_{Y|D}(1|d)$ (cf. Table II). Contrast this with Caucasian females (middle column), who have virtually no change in their recidivism rates since they are a priori close to the final ones (Table II). Another interesting observation is that middle aged Caucasian males with 1 to 3 prior counts see an increase in percentage recidivism. This is most likely an indirect effect of changes to the features X rather than a direct increase. One such source of increase is the age-increasing mapping shown in Fig. 4 (right) from reoffending young Caucasian males with a felony and 1 to 3 priors.

TABLE II
DEPENDENCE OF THE OUTCOME VARIABLE ON THE DISCRIMINATION VARIABLE BEFORE AND AFTER THE PROPOSED TRANSFORMATION. F AND M INDICATE FEMALE AND MALE, AND A-A AND C INDICATE AFRICAN-AMERICAN AND CAUCASIAN

D	Before transformation		After transformation	
(gender, race)	$p_{Y\|D}(0\|d)$	$p_{Y\|D}(1\|d)$	$p_{\hat{Y}\|D}(0\|d)$	$p_{\hat{Y}\|D}(1\|d)$
F, A-A	0.607	0.393	0.607	0.393
F, C	0.633	0.367	0.633	0.367
M, A-A	0.407	0.593	0.596	0.404
M, C	0.570	0.430	0.596	0.404

F. Pre-Processing Output for UCI Adult Data

We now look at a pre-processed version of the UCI Adult dataset. For this result, total variation distance was used to measure utility, and for a change, the combination of (1) and (3) were used to control discrimination, where we choose $p_{Y_T} = p_Y$ and $\epsilon_{y,d} = \epsilon = 0.15$. The distortion constraints remain as in Section IV-B. The corresponding optimal utility (total variation) was 0.014.

The result of applying (again a single realization) the mapping $p_{\hat{X},\hat{Y}|X,Y,D}$ to the data is given in Fig. 6, where we show percentages of high income individuals as a function of age and education before the transformation and percentage changes afterward. The original age and education (X) are plotted throughout Fig. 6 for ease of comparison. Note that changes in individual percentages may be larger than a factor of $1 \pm \epsilon$ because

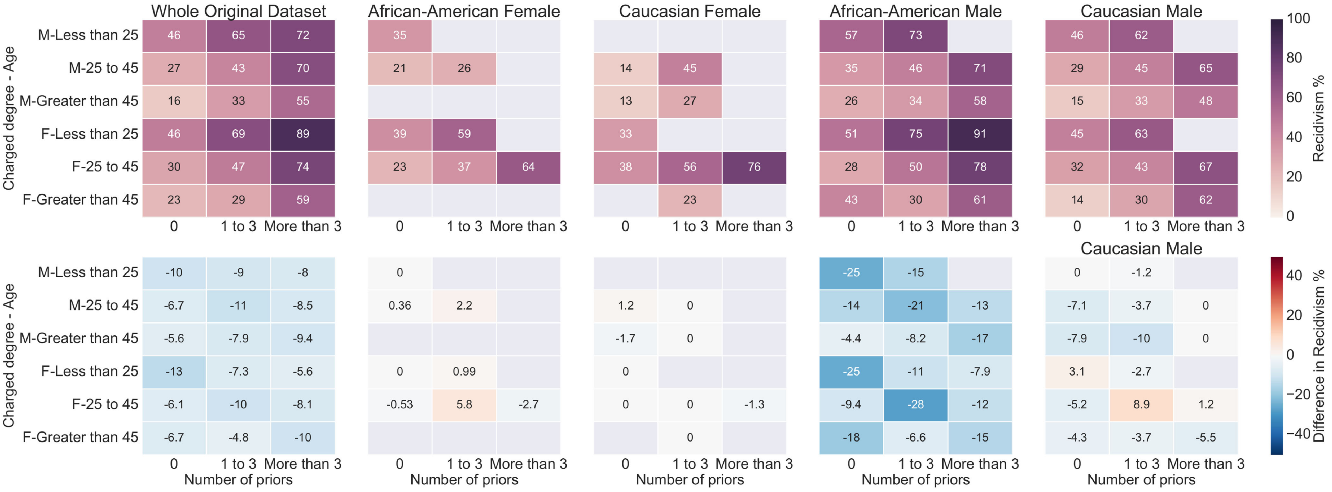

Fig. 5. Top row: Percentage recidivism rates in the original dataset as a function of charge degree, age and prior counts for the overall population (i.e., $p_{Y|X}(1|x)$) and for different groups ($p_{Y|X,D}(1|x,d)$). Bottom row: Change in percentages due to transformation, i.e., $p_{\hat{Y}|\hat{X},D}(1|x,d) - p_{Y|X,D}(1|x,d)$, etc. Values for cohorts of charge degree, age, and prior counts with fewer than 20 samples are not shown. The discrimination and distortion constraints are set to $\epsilon = 0.1$ and $c = 0.5$ respectively.

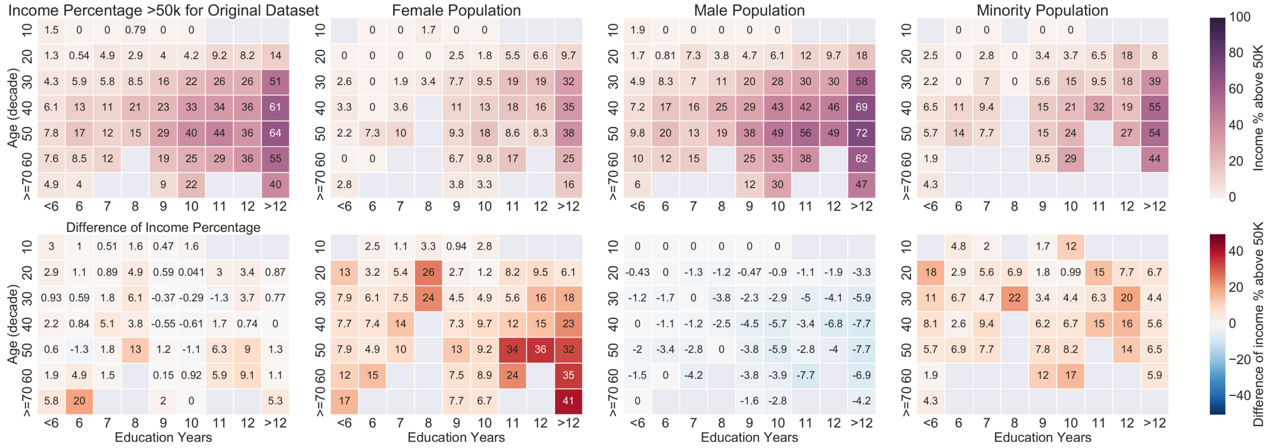

Fig. 6. Top row: High income percentages in the original Adult dataset as a function of age and education for the overall population (i.e., $p_{Y|X}(1|x)$) and for different groups $p_{Y|X,D}(1|x,d)$). Bottom row: Change in percentages due to transformation, i.e., $p_{\hat{Y}|\hat{X},D}(1|x,d) - p_{Y|X,D}(1|x,d)$, etc. Age-education pairs with fewer than 20 samples are not shown.

discrimination is not controlled by (1) at the level of age-education cohorts. The top left panel indicates that income is higher for more educated and middle-aged people, as expected. The second column shows that high income percentages are significantly lower for females and are accordingly increased by the transformation, most strongly for educated older women and younger women with only 8 years of education, and less so for other younger women. Conversely, the percentages are decreased for males but by much smaller magnitudes. Minorities receive small percentage increases but less than for women, in part because they are a more heterogeneous group consisting of both genders.

V. Conclusions

We proposed a flexible, data-driven optimization framework for probabilistically transforming data in order to reduce algorithmic discrimination, and applied it to two datasets. When used to train standard classifiers, the transformed datasets led to fairer classifications when compared to the original datasets. In the tested datasets, the reduction in discrimination comes with a small accuracy penalty due to the restrictions imposed on the pre-processing mapping. Moreover, our method is competitive with others in the literature, with the added benefits of enabling more explicit and precise control of both group and individual fairness as well as the possibility of multivariate, non-binary protected variables.

The differences between the original and transformed datasets revealed interesting discrimination patterns, as well as corrective adjustments for controlling discrimination while preserving utility of the data. Despite being programmatically generated, the optimized transformation satisfied properties that are sensible from a socio-demographic standpoint. The pre-processing transformation changed relationships between variables *within the datasets* (e.g., reducing recidivism risk for males who are African American in the recidivism dataset, increasing income for well-educated females in the UCI adult dataset) so that subsequently learned classifiers will then reflect these changes.

The flexibility of the approach allows numerous extensions using different measures and constraints for utility preservation, discrimination, and individual distortion control. Investigating such extensions, developing additional theoretical characterizations of the proposed framework, and quantifying the impact of the transformations and non-i.i.d. samples on additional supervised learning tasks will be pursued in future work.

Appendix A
Proof of Proposition 1

Proof: Considering first the objective function, the distribution $p_{X,Y}$ is a given quantity while

$$p_{\hat{X},\hat{Y}}(\hat{x},\hat{y}) = \sum_{d,x,y} p_{D,X,Y}(d,x,y) p_{\hat{X},\hat{Y}|D,X,Y}(\hat{x},\hat{y}|d,x,y)$$

is seen to be a linear function of the mapping $p_{\hat{X},\hat{Y}|D,X,Y}$, i.e., the optimization variable. Hence if the statistical dissimilarity $\Delta(\cdot,\cdot)$ is convex in its first argument with the second fixed, then $\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$ is a convex function of $p_{\hat{X},\hat{Y}|D,X,Y}$ by the affine composition property [41]. This condition is satisfied for example by all f-divergences [42], which are jointly convex in both arguments, and by all Bregman divergences [43]. If instead $\Delta(\cdot,\cdot)$ is only quasiconvex in its first argument, a similar composition property implies that $\Delta(p_{\hat{X},\hat{Y}}, p_{X,Y})$ is a quasiconvex function of $p_{\hat{X},\hat{Y}|D,X,Y}$ [41].

For discrimination constraint (1), the target distribution p_{Y_T} is assumed to be given. The conditional distribution $p_{\hat{Y}|D}$ can be related to $p_{\hat{X},\hat{Y}|D,X,Y}$ as follows:

$$p_{\hat{Y}|D}(\hat{y}|d) = \sum_{\hat{x}} \sum_{x,y} p_{X,Y|D}(x,y|d) p_{\hat{X},\hat{Y}|D,X,Y}(\hat{x},\hat{y}|d,x,y).$$

Since $p_{X,Y|D}$ is given, $p_{\hat{Y}|D}$ is a linear function of $p_{\hat{X},\hat{Y}|D,X,Y}$. Hence by the same composition property as above, (1) is a convex constraint, i.e., specifies a convex set, if the distance function $J(\cdot,\cdot)$ is quasiconvex in its first argument.

If constraint (2) is used instead of (1), then both arguments of J are linear functions of $p_{\hat{X},\hat{Y}|D,X,Y}$. Hence (2) is convex if J is jointly quasiconvex in both arguments.

Lastly, the distortion constraint (4) can be expanded explicitly in terms of $p_{\hat{X},\hat{Y}|D,X,Y}$ to yield

$$\sum_{\hat{x},\hat{y}} p_{\hat{X},\hat{Y}|D,X,Y}(\hat{x},\hat{y}|d,x,y) \delta\big((x,y),(\hat{x},\hat{y})\big) \le c_{d,x,y}.$$

Thus (4) is a linear constraint in $p_{\hat{X},\hat{Y}|D,X,Y}$ regardless of the choice of distortion metric δ. ■

Appendix B
Proof of Proposition 2

Proof: We will make use of the following result that follows directly from [44, Theorem 2.1]: for $m \triangleq |\mathcal{X}||\mathcal{Y}||\mathcal{D}|$,

$$\Pr\left(\|q_{X,Y,D} - p_{X,Y,D}\|_1 > \delta\right) \le 2^m \exp\left(-\frac{n\delta^2}{2}\right). \quad (18)$$

Assume

$$\|p_{D,X,Y} - q_{D,X,Y}\|_1 \le \tau. \quad (19)$$

Then the Data Processing Inequality for total variation [45] yields

$$\|p_{\hat{Y},D} - q_{\hat{Y},D}\|_1 \le \tau \quad (20)$$

and, equivalently

$$\|p_D - q_D\|_1 \le \tau. \quad (21)$$

Consequently, for all $y \in \mathcal{Y}$, $d \in \mathcal{D}$

$$\begin{aligned}
\tau &\ge p_{\hat{Y},D}(y,d) - q_{\hat{Y},D}(y,d) && (22)\\
&= p_D(d) p_{\hat{Y}|D}(y|d) - q_D(d) q_{\hat{Y}|D}(y|d) && (23)\\
&\ge p_D(d) p_{\hat{Y}|D}(y|d) - (p_D(d) + \tau) q_{\hat{Y}|D}(y|d) && (24)\\
&\ge p_D(d)(p_{\hat{Y}|D}(y|d) - q_{\hat{Y}|D}(y|d)) - \tau, && (25)
\end{aligned}$$

where the first and second inequalities follow from (20) and (21), respectively. Thus, if (19) holds, then

$$p_{\hat{Y}|D}(y|d) - q_{\hat{Y}|D}(y|d) \le \frac{2\tau}{p_D(d)}. \quad (26)$$

An equivalent procedure can be used to lower bound the left-hand side of the previous equation, resulting in

$$\left| p_{\hat{Y}|D}(y|d) - q_{\hat{Y}|D}(y|d) \right| \le \frac{2\tau}{p_D(d)} \; \forall y \in \mathcal{Y}, d \in \mathcal{D}. \quad (27)$$

Further assuming that $J(p_{\hat{Y}|D}(y|d), p_{Y_T}(y)) \le \epsilon$, a direct application of the triangle inequality produces

$$J\left(q_{\hat{Y}|D}(y|d), p_{Y_T}(y)\right) \le \epsilon + \frac{2\tau}{p_{Y_T}(y) p_D(d)} \quad (28)$$

Choosing $\tau = \sqrt{\frac{2}{n}(\ln \frac{1}{\beta} + m)}$, and combining (19), (28), and (18), we have that with probability $1-\beta$

$$J\Big(q_{\hat{Y}|D}(y|d), p_{Y_T}(y)\Big) \le \epsilon + \frac{2\sqrt{2}}{\sqrt{n} p_{Y_T}(y) p_D(d)} \left(\sqrt{\ln \frac{1}{\beta} + m}\right). \quad (29)$$

The results follows from the assumption that $p_{Y_T}(y) p_D(d) > 0$.

For the second claim, we start by applying the triangle inequality:

$$\begin{aligned}
\Delta\left(q_{X,Y}, q_{\hat{X},\hat{Y}}\right) \le\ & \Delta\left(p_{X,Y}, p_{\hat{X},\hat{Y}}\right) + \Delta\left(q_{X,Y}, p_{X,Y}\right)\\
& + \Delta\left(q_{\hat{X},\hat{Y}}, p_{\hat{X},\hat{Y}}\right)\\
\le\ & \mu + \Delta\left(q_{X,Y}, p_{X,Y}\right)\\
& + \Delta\left(q_{\hat{X},\hat{Y}}, p_{\hat{X},\hat{Y}}\right)\\
\le\ & \mu + 2\Delta\left(q_{X,Y}, p_{X,Y}\right), \qquad (30)
\end{aligned}$$

where the last inequality follows from the Data Processing Inequality for total variation [45]. Applying (18) and defining $m \triangleq |\mathcal{X}||\mathcal{Y}|$, we have for $\tau \ge 0$

$$\Pr\left(\Delta\left(q_{X,Y}, p_{X,Y}\right) > \tau\right) \le \exp\left(-\frac{n\tau^2}{2} + m\right). \quad (31)$$

Letting $\tau = \sqrt{\frac{2}{n}(\ln\frac{1}{\beta} + m)}$ and combining (30) and (31), we have that with probability at least $1 - \beta$

$$\Delta\left(q_{X,Y}, q_{\hat{X},\hat{Y}}\right) \leq \mu + \frac{2\sqrt{2}}{\sqrt{n}}\left(\sqrt{\ln\frac{1}{\beta} + m}\right), \tag{32}$$

proving the second part of the proposition. ■

Appendix C
Proof of Proposition 3

Proof: We show the contrapositive of the statement of the proposition. Assume that

$$\left|\frac{p_{Y|D}(y|d)}{p_{Y_T}(y)} - 1\right| \leq \epsilon \, \forall y \in \mathcal{Y}, d \in \mathcal{D}. \tag{33}$$

Then

$$\begin{aligned} P_c(D|Y) &= \sum_{y \in \mathcal{Y}} \max_{d \in \mathcal{D}} p_{D|Y}(d|y) p_Y(y) \\ &= \sum_{y \in \mathcal{Y}} \max_{d \in \mathcal{D}} p_{Y|D}(y|d) p_D(d) \\ &\leq \sum_{y \in \mathcal{Y}} \max_{d \in \mathcal{D}} (1 + \epsilon) p_{Y_T}(y) p_D(d) \\ &= (1 + \epsilon) \max_{d \in \mathcal{D}} p_D(d), \end{aligned}$$

where the inequality follows by noting that (33) implies $p_{Y|D}(y|d) \leq (1 + \epsilon) p_{Y_T}(y)$ for all $y \in \mathcal{Y}$, $d \in \mathcal{D}$. Rearranging the terms of the last equality, we arrive at

$$\frac{P_c(D|Y)}{\max_{d \in \mathcal{D}} p_D(d)} \leq 1 + \epsilon,$$

and the result follows by observing that the left-hand side is the definition of $\mathsf{Adv}(D|Y)$. ■

References

[1] H. F. Ladd, "Evidence on discrimination in mortgage lending," *J. Econ. Perspectives*, vol. 12, no. 2, pp. 41–62, 1998.

[2] T. Calders and I. Žliobaitė, "Why unbiased computational processes can lead to discriminative decision procedures," in *Discrimination and Privacy in the Information Society*. New York, NY, USA: Springer, 2013, pp. 43–57.

[3] D. Pedreschi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 560–568.

[4] S. Hajian, "Simultaneous discrimination prevention and privacy protection in data publishing and mining," Ph.D. dissertation, Univ. Rovira i Virgili, 2013. [Online]. Available: https://arxiv.org/abs/1306.6805.

[5] B. Fish, J. Kun, and Á. D. Lelkes, "A confidence-based approach for balancing fairness and accuracy," in *Proc. SIAM Int. Conf. Data Mining*, 2016, pp. 144–152.

[6] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment," in *Proc. 26th Int. World Wide Web Conf.*, 2017, pp. 1171–1180.

[7] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware learning through regularization approach," in *Proc. IEEE 11th Int. Conf. Data Mining Workshops*, 2011, pp. 643–650.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3315–3323.

[9] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 259–268.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.*, 2012, pp. 214–226.

[11] Z. Zhang and D. B. Neill, "Identifying significant predictive bias in classifiers," in *Proc. NIPS Workshop Interpretable Mach. Learn. Complex Syst.*, 2016. [Online]. Available: https://arxiv.org/abs/1611.08292

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in *Proc. Innov. Theor. Comput. Sci. Conf.*, 2017.

[13] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, pp. 153–163, 2017.

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

[15] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 797–806.

[16] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf.*, vol. 33, no. 1, pp. 1–33, 2012.

[17] S. Hajian and J. Domingo-Ferrer, "A methodology for direct and indirect discrimination prevention in data mining," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 7, pp. 1445–1459, Jul. 2013.

[18] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *Proc. IEEE 23rd Int. Conf. Data Eng.*, 2007, pp. 106–115.

[19] S. Ruggieri, "Using t-closeness anonymity to control for non-discrimination," *Trans. Data Privacy*, vol. 7, no. 2, pp. 99–129, 2014.

[20] R. Zemel, Y. L. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 325–333.

[21] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 6, pp. 838–852, Jun. 2013.

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

[23] S. Salamatian *et al.*, "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," *IEEE Global Conf. Signal Inf. Process.*, 2013, pp. 269–272.

[24] F. P. Calmon, A. Makhdoumi, and M. Médard, "Fundamental limits of perfect privacy," in *Proc. In. Symp. Inf. Theory*, 2015, pp. 1796–1800.

[25] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, Jul. 2006.

[26] R. G. Gallager, *Information Theory and Reliable Communication*. New York, NY, USA: Wiley, 1968.

[27] ProPublica, "COMPAS Recidivism risk score data and analysis," 2017. [Online]. Available: https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-d ata-and-analysis

[28] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

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

[30] D. Pedreschi, S. Ruggieri, and F. Turini, "A study of top-k measures for discrimination discovery," in *Proc. ACM Symp. Appl. Comput.*, 2012, pp. 126–131.

[31] I. Žliobaitė, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *Proc. IEEE Int. Conf. Data Mining*, 2011, pp. 992–1001.

[32] J. Pearl, "Comment: Understanding Simpson's paradox," *Amer. Statist.*, vol. 68, no. 1, pp. 8–13, 2014.

[33] K. D. Johnson, D. P. Foster, and R. A. Stine, "Impartial predictive modeling: Ensuring fairness in arbitrary models," 2016, arXiv:1608.00528.

[34] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances Neural Inform. Process. Syst.*, pp. 4066–4076, Dec. 2017.

[35] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances Neural Inform. Process. Syst.*, pp. 656–666, Dec. 2017.

[36] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3/4, pp. 231–357, 2015.

[37] ProPublica, "Machine bias," 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-crimina l-sentencing

[38] Northpointe, Inc., "COMPAS—The most scientifically advanced risk and needs assessments." 2017. [Online]. Available: http://www.northpointeinc.com/risk-needs-assessment

[39] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 83, pp. 1–5, 2016.

[40] E. Jones *et al.*, "SciPy: Open source scientific tools for Python," 2001. [Online]. Available: http://www.scipy.org/

[41] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[42] I. Csiszár and P. C. Shields, "Information theory and statistics: A tutorial," *Found. Trends Commun. Inf. Theory*, vol. 1, no. 4, pp. 417–528, 2004.

[43] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *J. Mach. Learn. Res.*, vol. 6, pp. 1705–1749, 2005.

[44] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger, "Inequalities for the l1 deviation of the empirical distribution," Hewlett-Packard Labs, Palo Alto, CA, USA, Tech. Rep. HPL-2003-97R1, 2003.

[45] F. Liese and I. Vajda, "On divergences and informations in statistics and information theory," *IEEE Trans. Inf. Theory*, vol. 52, no. 10, pp. 4394–4412, Oct. 2006.

Flavio du Pin Calmon (S'11–M'17) received the B.Sc. degree in communications engineering from the Universidade de Brasilia, Brasília, Brazil, the M.Sc. degree in electrical engineering from the Universidade Estadual de Campinas, Campinas, Brazil, and Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, MA, USA. He is an Assistant Professor with the John A. Paulson School of Engineering and Applied Sciences, Harvard University. Prior to joining Harvard, he was the inaugural Social Good Postdoctoral Fellow with the IBM T. J. Watson Research Center. His research interests include information theory, data analytics, machine learning, fairness, security, and privacy.

Dennis Wei (S'09–M'11) received S.B. degrees in electrical engineering and in physics in 2006, the M.Eng. degree in electrical engineering in 2007, and the Ph.D. degree in electrical engineering in 2011, all from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

He is a Research Staff Member with IBM Research, Thomas J. Watson Research Center, Yorktown Heights, NY, USA. From 2011 to 2013, he was a Postdoctoral Research Fellow with the Electrical Engineering and Computer Science Department, University of Michigan, Ann Arbor, MI, USA. His research interests include signal processing, machine learning, statistics, optimization, algorithmic fairness, interpretability of machine learning models, privacy-preserving data release, graphical models, adaptive sampling, and sparse filter design.

Dr. Wei received a Best Paper Honorable Mention at the 2015 SIAM International Conference on Data Mining (2015), a Notable Paper Award at the 2013 International Conference on Artificial Intelligence and Statistics, and coauthored a Best Student Paper at the 2013 IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing. He is an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING.

Bhanukiran Vinzamuri (M'17) received the B.Tech. and masters's degrees in computer science from the International Institute of Information Technology Hyderabad, Hyderabad, India, and the Ph.D. degree in computer science from Wayne State University, Detroit, MI, USA. He is a research staff member with IBM Research, Yorktown Heights, NY, USA. His primary research interests include sparse learning, nonconvex optimization, and causal inference.

Karthikeyan Natesan Ramamurthy received the Ph.D. degree in electrical engineering from Arizona State University, Tempe, AZ, USA. He is a Research Staff Member with IBM Research AI, Thomas J. Watson Research Center, Yorktown Heights, NY, USA. His broad research interests are in understanding the geometry and topology of high-dimensional data and developing theory and methods for efficiently modeling the data. He has also been intrigued by the interplay between humans, machines, and data, and the societal implications of machine learning. He is the recipient of best paper awards at the 2015 IEEE International Conference on Data Science and Advanced Analytics and the 2015 SIAM International Conference on Data Mining. He is an Associate Editor for the *Digital Signal Processing* journal.

Kush R. Varshney (S'00–M'10–SM'15) was born in Syracuse, NY, USA, in 1982. He received the B.S. degree (*magna cum laude*) in electrical and computer engineering with honors from Cornell University, Ithaca, NY, USA, in 2004, and the S.M. degree in 2006 and the Ph.D. degree in 2010, both in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.

While at MIT, he was a National Science Foundation Graduate Research Fellow. He is a principal research staff member and manager with IBM Research AI, Thomas J. Watson Research Center, Yorktown Heights, NY, USA, where he leads the Learning and Decision Making Group. He is the founding Co-Director of the IBM Science for Social Good initiative. He applies data science and predictive analytics to human capital management, healthcare, olfaction, computational creativity, public affairs, international development, and algorithmic fairness, which has led to recognitions such as the 2013 Gerstner Award for Client Excellence for contributions to the WellPoint team and the Extraordinary IBM Research Technical Accomplishment for contributions to workforce innovation and enterprise transformation. He conducts academic research on the theory and methods of statistical signal processing and machine learning. He is the recipient of best paper awards at the Fusion 2009, SOLI 2013, KDD 2014, and SDM 2015 conferences. He is a member of the Partnership on AI's Safety-Critical AI working group.