entropy

MDPI

Article

A Maximal Correlation Framework for Fair Machine Learning

Joshua Lee [1,†,‡], Yuheng Bu [1,*,‡], Prasanna Sattigeri [2], Rameswar Panda [2], Gregory W. Wornell [1], Leonid Karlinsky [2] and Rogerio Schmidt Feris [2]

[1] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; jk_lee@mit.edu (J.L.); gww@mit.edu (G.W.W.)

[2] MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA 02139, USA; psattig@us.ibm.com (P.S.); rpanda@ibm.com (R.P.); leonidka@il.ibm.com (L.K.); rsferis@us.ibm.com (R.S.F.)

* Correspondence: buyuheng@mit.edu

† Current address: Snap Inc., Santa Monica, CA 90405, USA.

‡ These authors contributed equally to this work.

Abstract: As machine learning algorithms grow in popularity and diversify to many industries, ethical and legal concerns regarding their fairness have become increasingly relevant. We explore the problem of algorithmic fairness, taking an information–theoretic view. The maximal correlation framework is introduced for expressing fairness constraints and is shown to be capable of being used to derive regularizers that enforce independence and separation-based fairness criteria, which admit optimization algorithms for both discrete and continuous variables that are more computationally efficient than existing algorithms. We show that these algorithms provide smooth performance–fairness tradeoff curves and perform competitively with state-of-the-art methods on both discrete datasets (COMPAS, Adult) and continuous datasets (Communities and Crimes).

Keywords: fairness; HGR maximal correlation; independence criterion; separation criterion

check for updates

Citation: Lee, J.; Bu, Y.; Sattigeri, P.; Panda, R.; Wornell, G.W.; Karlinsky, L.; Schmidt Feris, R. A Maximal Correlation Framework for Fair Machine Learning. *Entropy* **2022**, *24*, 461. https://doi.org/10.3390/e24040461

Academic Editor: Friedhelm Schwenker

Received: 15 February 2022
Accepted: 24 March 2022
Published: 26 March 2022

1. Introduction

The use of machine learning in many industries has raised many ethical and legal concerns, especially that of fairness and bias in predictions, e.g., [1,2]. As systems are trusted to aid or make decisions regarding loan applications, criminal sentencing, and even health care, it is vital that unfair biases do not influence them.

However, mitigating these biases is complicated by ever-changing perspectives on fairness, and a good system for enforcing fairness must be adaptable to new settings. In particular, there are often competing notions on fairness. Two of these popular notions are independence and separation (a third condition, sufficiency, is beyond the scope of this paper), as discussed in [3]. Independence ensures that predictions are independent from membership in a protected class, so that one achieves equal favorable outcome rates across all groups, and it arises in applications such as affirmative action [4]. Separation is designed to achieve equal type I/II error rates across all groups by enforcing independence between predictions and membership in a protected class conditional on the class label. This criterion is used to measure fairness in recidivism predictions and bank loan applications. A significant body of work, including [3,5–7], has gone into explaining that independence and separation are inherently incompatible for non-trivial cases, and their applicability needs to be determined by the application and the stakeholders. This motivates us to construct a framework that is flexible enough to handle different fairness criteria and to do it with different modalities of data (discrete vs. continuous data, for example).

This bias mitigation must also be balanced out with the system's usefulness, and often, one must tune the tradeoff between the fairness (as measured in the particular context) and performance according to a current situation, which can be a difficult process if the tradeoff curve is not smooth. Generating the frontier of possible values can be computationally infeasible or impossible if the algorithm does not have a regularization parameter to adjust (see, [8,9]), thus making it difficult to achieve this balance, which makes the fast generation of fair classifiers even more important.

Different contexts also require different points of intervention during the learning process to ensure fairness. *Pre-processing* approaches ([8,10–14]) modify the data to eliminate bias, whereas *post-processing* approaches ([15–18]) modify learned features/predictions from existing models to be more fair. We focus on the *in-processing* approach [9,19–21], where the fairness criteria are directly incorporated in the training objective to produce fairer learned features. Motivated by few-shot applications where only a pre-trained network and few samples labeled with the sensitive attribute are available, we also seek a method that is applicable in a post-processing manner when we have access to only a small number of samples labeled with the sensitive attribute that we wish to be fair about, which would arise in settings where collecting this information can be very difficult.

In this paper, we frame the ideas of independence and separation in a way that allows a relevant regularizer or penalty term to be derived in addition to a measure of fairness, which is useful in enforcing fairness while also tractable, admitting an optimization algorithm (e.g., if used as an objective for a neural net trained using gradient descent, it must be differentiable), and easily computed. Existing approaches can struggle with efficiency, can fail to provide good control over the performance–fairness tradeoff, and/or can only deal with either discrete or continuous data.

We make the following contributions in this paper:

- We present a universal framework justified by an information–theoretic view that can inherently handle the popular fairness criteria, namely independence and separation, while seamlessly adopting both discrete and continuous cases, which uses the maximal correlation to construct measures of fairness associated with different criteria; then, we use these measures to further develop fair learning algorithms in a fast, efficient, and effective manner.
- We show empirically that these algorithms can provide the desired smooth tradeoff curve between the performance and the measures of fairness on several standard datasets (COMPAS, Adult, and Communities and Crimes), so that a desired level of fairness can be achieved.
- Finally, we perform experiments to illustrate that our algorithms can be used to impose fairness on a model originally trained without any fairness constraint in the few-shot regime, which further demonstrates the versatility of our algorithms in a post-processing setup.

2. Background

2.1. Fairness Objectives in Machine Learning

Consider the standard supervised learning scenario where we predict the value of a target variable $Y \in \mathcal{Y}$ using a set of decision or predictive variables $X \in \mathcal{X}$ with training samples $\{(x_1, y_1), \dots, (x_n, y_n)\}$. For example, X may be information about an individual's credit history, and Y is whether the individual will pay back a certain loan. In general, we wish to find features $f(x)$, which are predictive of Y, so that we can construct a good predictor $\hat{y} = T(f(x))$ of y under some loss criteria $L(\hat{y}, y)$.

Now, suppose we have some sensitive attributes $D \in \mathcal{D}$ we wish to be "fair" about (e.g., race, gender), and training samples $\{(x_1, y_1, d_1), \dots, (x_n, y_n, d_n)\}$. For example, in the criminal justice system, predictions about the chance of recidivism of a convicted criminal (Y) given factors such as the nature of the crime and the number of prior arrests (X) should not be determined by race (D). This is a known issue with the COMPAS recidivism score, which, despite not using race as an input to make decisions, still leads to systematic bias toward members of certain races in the output score as in [22,23].

The two most popular criteria for fairness are independence and separation. Independence states that for a feature to be fair, it must satisfy the independence property $\hat{Y} \perp D$ or $f(x) \perp D$. The intuition is simple: if the prediction/feature is independent of the sensitive attribute, then no information about the sensitive attribute is used to predict Y. This criterion has been studied under the lens of *demographic parity* and *disparate impact* in [3], and it admits a class of fairness measures based on the degree of dependence between $f(X)$ and D. For example, independence is satisfied if and only if the mutual information $I(f(X);D)$ is zero. When D is binary, another popular class of measures used by the US Equal Employment Opportunity Commission [4] is the disparate impact, which is defined as $\mathbb{D}\big(\mathbb{P}(Y|D=1);\mathbb{P}(Y|D=0)\big) = \frac{\mathbb{P}(\hat{Y}=1|D=0)}{\mathbb{P}(\hat{Y}=1|D=1)}$.

Separation requires the conditional independence property $(\hat{Y} \perp D)|Y$ or $(f(X) \perp D)|Y$. This criterion allows for a violation of demographic parity to the extent that it is justified by the target variable. In the general case, this criterion suggests a fairness measure based on the conditional dependence between $\hat{Y}$ and D conditioned on Y. In the case where D is binary, we obtain the *equalized opportunities* (EO) measures in [3], which are given by the differences in error rates for the two groups (e.g., the difference between the false positive rates for $D = 0, 1$). For a more complete discussion of the advantages and disadvantages of these two criteria, please refer to [3].

2.2. Maximal Correlation

Since these fairness criteria are expressed as enforcing independencies with respect to joint distributions, we look for constraints that reduce the dependency between variables. In particular, the right formulation of correlation between learned features and sensitive attributes can provide a framework for measuring and optimizing for fairness. One effective measure applicable to both continuous and discrete data is the Hirschfeld–Gebelein–Renyi (HGR) maximal correlation, which is a measure of nonlinear correlation that originated in [24] and is further developed in [25,26]. The HGR maximal correlation between two random variables is equal to zero if and only if the two variables are independent, and it increases in value the more correlated they are (i.e., the more biased/unfair).

Definition 1. *For two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, given $1 \le k \le K-1$ with $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, the HGR maximal correlation problem is*

$$(\mathbf{f}^*, \mathbf{g}^*) \triangleq \underset{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^k,\ \mathbf{g}\colon \mathcal{Y}\to\mathbb{R}^k}{\arg\max}\ \mathbb{E}\Big[\mathbf{f}^{\mathrm{T}}(X)\,\mathbf{g}(Y)\Big], \tag{1}$$

with constraints

$$\mathbb{E}[\mathbf{f}(X)] = \mathbb{E}[\mathbf{g}(Y)] = \mathbf{0}, \quad \mathbb{E}\Big[\mathbf{f}(X)\mathbf{f}^{\mathrm{T}}(X)\Big] = \mathbb{E}\Big[\mathbf{g}(Y)\mathbf{g}^{\mathrm{T}}(Y)\Big] = \mathbf{I}, \tag{2}$$

and expectations taken over $P_{X,Y}$. We refer to $\mathbf{f}^*$ and $\mathbf{g}^*$ as maximal correlation functions, with $\mathbf{f}^* = (f_1^*, \ldots, f_k^*)^{\mathrm{T}}$ and $\mathbf{g}^* = (g_1^*, \ldots, g_k^*)^{\mathrm{T}}$, and the associated maximal correlations are

$$\sigma(f_i^*\, g_i^*) \triangleq \mathbb{E}[f_i^*(X)\, g_i^*(Y)],\ \textit{for}\ \ i = 1, \ldots, k, \tag{3}$$

and the HGR maximal correlation is

$$\mathrm{HGR}_k(X,Y) \triangleq \mathbb{E}\Big[\mathbf{f}^{*\mathrm{T}}(X)\,\mathbf{g}^*(Y)\Big] = \sum_{i=1}^{k} \sigma(f_i^*\, g_i^*). \tag{4}$$

Note that the original definition of HGR maximal correlation is the special case of our definition when $k = 1$ (see, [27]). This generalization of maximal correlation analysis enables us to produce more than one feature mapping by solving the maximal correlation problem, and these feature mappings can be used in other applications, including ensemble learning, multi-task learning, and transfer learning [28,29].

2.3. Related Work

Independence and separation have been studied in many works. Most existing approaches fail to provide an efficient solution in both discrete/continuous settings. Ref. [11] develops an optimizer using absolute difference in odds $|\mathbb{P}(\hat{Y} = 1|D = 1) - \mathbb{P}(\hat{Y} = 1|D = 0)|$ as a regularizer, which requires discrete Y and D and was only applied to Naïve Bayes and Logistic Regression to enforce the independence criterion. In [16], a post-processing method is provided using a probabilistic combination of classifiers to achieve the desired ROC curves, which only applies when D is discrete. Alternatively, Ref. [8] proposes pre-processing the data beforehand to enforce fairness before learning, based on randomized mappings of the data subject to a fairness constraint defined by $J = \max(|\frac{\mathbb{P}(\hat{Y}=1|D=1)}{\mathbb{P}(\hat{Y}=1|D=0)} - 1|, |\frac{\mathbb{P}(\hat{Y}=1|D=0)}{\mathbb{P}(\hat{Y}=1|D=1)} - 1|)$. Again, this method is only designed for independence with discrete Y and D, and it requires processing the entire dataset, which is computationally complex. Ref. [30] propose the use of a robust log-loss predictor for fairness, but in practice, it requires that Y be discrete.

Other methods can also be limited in their ability to handle all dependencies between variables. Ref. [31] uses a covariance-based constraint to enforce fairness, so it likely would not do well on other metrics. Furthermore, it is strictly a linear penalty rather than our non-linear formulation and penalizes the predictions of the system rather than the features learned. This limits the relationships between variables it can capture. An adversarial method is proposed in [20] to enforce independence or separation, but it requires the training of an adversary to predict the sensitive attribute, which can introduce issues of convergence and bias.

Recently, Ref. [9] propose the use of the HGR maximal correlation as a regularizer for either the independence or the separation constraint. In contrast to our approach dealing with the maximal correlation directly, they use a χ^2 divergence computed over a mesh grid to upper bound the HGR maximal correlation during the optimization of the classifier (either a linear regressor or a Deep Neural Net (DNN)). This method applies to cases where X is continuous and Y and D are either continuous or discrete variables, but it scales poorly with the bandwidth and dimensionality of D, and it treats the discrete case in the same way as the continuous case, resulting in slow performance on discrete datasets.

There are other works that use either an HGR-based or mutual information-based formulation of fairness but do not generalize to more than one setting. Refs. [32,33] use correlation-based regularizers but can only be used in the independence case. Furthermore, Ref. [33] only works with discrete targets, and only uses a single mode of the HGR maximal correlation (as opposed to multiple modes, which our method makes use of) for regularization, which limits the information it can encapsulate, and it is also not designed for continuous sensitive attributes. Ref. [34] also develops a method that can only be used for independence, and it requires training an additional network in order to evaluate a bound for the mutual information which can be used to as a fairness penalty, thus increasing the complexity and required runtime. Finally, Ref. [35] approximates the mutual information with a variational formulation, but it does not include a formulation for continuous labels.

3. Maximal Correlation for Fairness

Equipped with the HGR maximal correlation as a measure of dependence, we explore its use as a fairness penalty. Depending on the data modality (discrete/continuous) and the fairness criteria (independence/separation), the resulting fair learning algorithm takes different specifically tailored forms. In this section, we demonstrate how to derive these regularizers and algorithms to ensure the aforementioned fairness objectives for both discrete and continuous cases.

3.1. Maximal Correlation for Discrete Learning

In this subsection, the decision variable X, target variable Y, and sensitive attribute D are discrete random variables defined on alphabets $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{D}$, respectively.

We first describe how to solve the discrete maximal correlation problem using a divergence transfer matrix (DTM)-based approach. As it is shown later, it is more convenient to work with their equivalent representation via DTM instead of the joint distribution $P_{X,Y}$.

Definition 2. *The divergence transfer matrix (DTM) $\mathbf{B}_{Y,X} \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{X}|}$ associated with joint distribution $P_{X,Y}$ is given by*

$$\mathbf{B}_{X,Y}(x,y) \triangleq \frac{P_{X,Y}(x,y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}. \tag{5}$$

The following useful result expresses that the maximal correlation problem can be solved by simply computing the singular value decomposition (SVD) of the DTM $\mathbf{B}$ in the discrete case.

Theorem 1 ([27]). *Assume that the SVD of DTM $\mathbf{B}_{Y,X}$ takes the form*

$$\mathbf{B}_{Y,X} = \sum_{i=0}^{K-1} \sigma_i \psi_i^Y (\psi_i^X)^{\mathrm{T}}, \tag{6}$$

with singular values $\sigma_0 \geq \sigma_1 \geq \cdots \geq \sigma_{K-1}$, singular vectors ψ_i^Y, ψ_i^X, and $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$. Then, we have

$$\sigma_0 = 1, \quad \psi_0^X(x) = \sqrt{P_X(x)}, \quad \psi_0^Y(y) = \sqrt{P_Y(y)}, \tag{7}$$

and the maximal correlation functions are related to the singular vectors in the SVD:

$$f_i^*(x) = \frac{\psi_i^X(x)}{\sqrt{P_X(x)}}, \quad g_i^*(x) = \frac{\psi_i^Y(y)}{\sqrt{P_Y(y)}}, \tag{8}$$

with associated maximal correlations $\sigma(f_i^*\, g_i^*) = \sigma_i$, for $i = 1, \cdots, K-1$. Thus, the conditional distribution $P_{Y|X}$ has the following decomposition:

$$P_{Y|X}(y|x) = P_Y(y)\Big[1 + \sum_{i=1}^{K-1} \sigma_i f_i^*(x) g_i^*(y)\Big]. \tag{9}$$

As we can see from this theorem, the singular values σ_i (since the associated maximal correlations is equal to the corresponding singular values of DTM, we abuse the notation a little bit and use σ to denote both of them) of the matrix $\mathbf{B}_{Y,X}$ essentially characterize the dependence between two discrete random variables, and the singular vectors $\Phi^X = [\psi_1^X, \cdots, \psi_k^X]$ and $\Phi^Y = [\psi_1^Y, \cdots, \psi_k^Y]$ are equivalent to the maximal correlation functions $\mathbf{f}$ and $\mathbf{g}$.

Since our goal is to construct feature mappings $\mathbf{f}(x)$ under fairness constraints, our algorithms in the discrete case are built on the following variational characterization of an SVD, which does not involve $\mathbf{g}(y)$:

Lemma 1 ([36]). *For any $k \leq K-1$ and $\Phi_X \in \mathbb{R}^{|\mathcal{X}|\times(k+1)}$,*

$$\max_{\Phi_X^{\mathrm{T}}\Phi_X = \mathbf{I}} \|\mathbf{B}\Phi_X\|_{\mathrm{F}}^2 = \sum_{i=0}^{k} \sigma_i^2, \tag{10}$$

where $\|A\|_{\mathrm{F}} \triangleq \sqrt{\mathrm{tr}(A^{\mathrm{T}}A)}$ denotes the Frobenius norm.

3.1.1. Independence

To ensure sufficient independence, we must construct feature mappings $\mathbf{f}: \mathcal{X} \to \mathbb{R}^k$ so that the maximal correlations between $\mathbf{f}(X)$ and Y are large, while the ones between

$\mathbf{f}(X)$ and D are small. Motivated by Lemma 1 and Theorem 1, we propose the following DTM-based approach to construct $\mathbf{f}$:

$$\max_{\Phi \in \mathbb{R}^{|\mathcal{X}| \times (k+1)}: \Phi^{\mathrm{T}}\Phi = \mathbf{I}} \|\mathbf{B}_{Y,X}\Phi\|_{\mathrm{F}}^2 - \lambda \|\mathbf{B}_{D,X}\Phi\|_{\mathrm{F}}^2, \tag{11}$$

where $\mathbf{B}_{Y,X}$ and $\mathbf{B}_{D,X}$ denote the DTMs of distribution $P_{Y,X}$ and $P_{D,X}$, respectively, and λ is the regularization coefficient that controls the penalty of the maximal correlations between $\mathbf{f}(X)$ and D. $\Phi^* = [\phi_0^*, \phi_1^*, \cdots, \phi_k^*]$ is the solution of the optimization problem (11). As shown in Theorem 1, $\mathbf{B}_{Y,X}$ and $\mathbf{B}_{D,X}$ have a shared right singular vector $\sqrt{P_X(x)}$, and we can let $\phi_0^* = \sqrt{P_X(x)}$. Then, the feature mappings for independence can be obtained by normalizing other column vectors in Φ^*

$$f_i(x) = \phi_i^*(x) / \sqrt{P_X(x)},\ i = 1, \cdots, k. \tag{12}$$

We have the following remarks:

(1) The optimization problem in (11) can be written as $\max \operatorname{tr}(\Phi^{\mathrm{T}} (\mathbf{B}_{Y,X}^{\mathrm{T}} \mathbf{B}_{Y,X} - \lambda \mathbf{B}_{D,X}^{\mathrm{T}} \mathbf{B}_{D,X}) \Phi)$, and it can be solved exactly by computing the eigen decomposition of $\mathbf{B}_{Y,X}^{\mathrm{T}} \mathbf{B}_{Y,X} - \lambda \mathbf{B}_{D,X}^{\mathrm{T}} \mathbf{B}_{D,X}$.

(2) Lemma 1 states that the Frobenius norm squared $\|\mathbf{B}_{Y,X} F\|_{\mathrm{F}}^2$ corresponds to the squared sum of the singular values. Actually, the following lemma shows that $\|\mathbf{B}_{Y,X} F\|_{\mathrm{F}}^2$ can be further related to the mutual information $I(X;Y)$ when the dependence between X and Y is weak.

Lemma 2 ([27])**.** *Let $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ be ϵ-dependent random variables; i.e., the χ^2-divergence is bounded $D_{\chi^2}(P_{X,Y} \| P_X P_Y) \leq \epsilon$, then*

$$I(X;Y) = \frac{1}{2} \sum_{i=1}^{K-1} \sigma_i^2 + o(\epsilon^2). \tag{13}$$

(3) As suggested by Lemma 2, the optimization problem in (11) can also be interpreted as maximizing the mutual information between $\mathbf{f}(X)$ and Y while penalizing the mutual information $I(\mathbf{f}(X); D)$.

Once we solve (11) and obtain the feature mappings $\mathbf{f}(x)$, we can obtain the corresponding maximal correlation function $\mathbf{g}(y)$ for the target variable Y via one step of the alternating conditional expectations algorithm by [37]:

$$g_i(y) \propto \mathbb{E}_{p_{X|Y}(\cdot|y)}[f_i(X)],\ i = 1, \ldots, k. \tag{14}$$

In turn, $\mathbf{g}(y)$ can be computed by further normalizing the conditional expectations of $\mathbf{f}(X)$, so that the condition $\mathbb{E}\left[\mathbf{g}(Y)\mathbf{g}^{\mathrm{T}}(Y)\right] = \mathbf{I}$ is satisfied. Finally, the predictions $\hat{Y}$ can be made following the Maximum A Posteriori (MAP) rule, where the posteriori distribution $P_{Y|X}(y|x)$ can be approximately computed by plugging the learned feature mappings $\mathbf{f}(X)$ and $\mathbf{g}(Y)$ into (9), i.e.,

$$\hat{Y} = \arg\max_{y \in \mathcal{Y}} P_Y(y) \left[1 + \sum_{i=1}^{k} \sigma_i f_i(x) g_i(y)\right]. \tag{15}$$

3.1.2. Separation

For the separation criterion, we want to ensure sufficient conditional independence $(f(X) \perp D)|Y$. Here, we cannot simply replace the $\mathbf{B}_{D,X}$ in (11) with a conditional DTM, as it involves three random variables and thus cannot be usefully expressed as a matrix. Since maximal correlation is related to mutual information as shown in Lemma 2, we consider the following formulation:

$$\begin{aligned}
&\max_{\mathbf{f}} I(\mathbf{f}(X);Y) - \lambda I(\mathbf{f}(X);D,Y) \\
&= \max_{\mathbf{f}} I(\mathbf{f}(X);Y) - \lambda\big(I(\mathbf{f}(X);Y) + I(\mathbf{f}(X);D|Y)\big) \\
&= \max_{\mathbf{f}}(1-\lambda)I(\mathbf{f}(X);Y) - \lambda I(\mathbf{f}(X);D|Y),
\end{aligned} \tag{16}$$

where the first equality follows from the chain rule of mutual information and $\lambda \in (0,1)$. Thus, we can control the conditional mutual information $I(\mathbf{f}(X);D|Y)$ by adding the joint mutual information $I(\mathbf{f}(X);D,Y)$ as a regularizer in the training process.

Note that Lemma 1 and Lemma 2 imply that mutual information can be approximated using DTM, as shown in (11) in an independence case. Accordingly, we approximate (16) using the following optimization problem to ensure the separation criterion for discrete data:

$$\max_{\Phi \in \mathbb{R}^{|\mathcal{X}| \times (k+1)}: \Phi^{\mathrm{T}}\Phi = \mathbf{I}} \|B_{Y,X}\Phi\|_{\mathrm{F}}^2 - \lambda \|B_{D\otimes Y,X}\Phi\|_{\mathrm{F}}^2, \tag{17}$$

where $D \otimes Y$ is the Cartesian product of D and Y, and $B_{D\otimes Y,X}$ denotes the DTM of distribution $P_{D\otimes Y,X}$. Once we obtained the solution Φ^*, we could follow similar steps as in the independence case to get $\mathbf{f}(x)$ and $\mathbf{g}(y)$ and make predictions for the test samples.

3.2. Maximal Correlation for Continuous Learning

When X, Y, and D are all continuous and real-valued, computing the HGR maximal correlation becomes much more difficult, since the space of functions over real numbers is not tractable. Thus, we turn to approximations and begin by limiting our scope of learning algorithms to those that train models (e.g., neural nets) via gradient descent (or SGD) using samples, which encompasses most of the commonly used methods. Then, it follows that any approximation of the HGR maximal correlation used must be differentiable to calculate the gradient. Thus, we restrict the space of maximal correlation functions to be the family of functions that can be learned by neural nets, allowing us to compute the gradient while still providing a rich set of functions to search over.

3.2.1. Independence

To ensure sufficient independence, we want to minimize the loss function $L(\hat{Y},Y)$ and the maximal correlation between $\mathbf{f}(X)$ and D. Then, our optimization (for a given λ) becomes:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} L(T(\mathbf{f}(X)),Y) + \lambda \mathrm{HGR}_k(\mathbf{f}(X),D), \tag{18}$$

where $\mathrm{HGR}_k(\mathbf{f}(X),D) = \max_{\mathbf{g},\,\mathbf{h}} \mathbb{E}\big[\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\,\mathbf{h}(D)\big]$, with $\mathbb{E}[\mathbf{g}(\mathbf{f}(X))] = \mathbb{E}[\mathbf{h}(D)] = \mathbf{0}$, and $\mathbb{E}\big[\mathbf{g}(\mathbf{f}(X))\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\big] = \mathbb{E}\big[\mathbf{h}(D)\mathbf{h}^{\mathrm{T}}(D)\big] = \mathbf{I}$. m is the dimension of the features $\mathbf{f}(X)$, k is the number of maximal correlation functions, and $\mathbf{g}\colon \mathbb{R}^m \to \mathbb{R}^k$, $\mathbf{h}\colon \mathcal{D} \to \mathbb{R}^k$ are the maximal correlation functions relating $\mathbf{f}(X)$ with D. Given the difficulty of enforcing the orthogonalization constraint, we use a variational characterization of the HGR maximal correlation called Soft-HGR proposed in [29], which relaxes the orthogonal constraint:

$$\mathrm{HGR}_{\mathrm{soft}}(X,Y) \triangleq \max_{\substack{\mathbb{E}[\mathbf{g}(X)]=\mathbf{0} \\ \mathbb{E}[\mathbf{h}(Y)]=\mathbf{0}}} \mathbb{E}\Big[\mathbf{g}^{\mathrm{T}}(X)\,\mathbf{h}(Y)\Big] - \frac{1}{2}\operatorname{tr}(\operatorname{cov}[\mathbf{g}(X)]\operatorname{cov}[\mathbf{h}(Y)]), \tag{19}$$

where $\mathrm{cov}[X]$ is the covariance matrix of X. [29] shows that this Soft-HGR formulation can be viewed as a low-rank approximation of the original HGR maximal correlation problem in the discrete case. Then, our learning objective becomes:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} \quad \max_{\substack{\mathbf{g}\colon \mathbb{R}^m\to\mathbb{R}^k,\ \mathbf{h}\colon \mathcal{D}\to\mathbb{R}^k \\ \mathbb{E}[\mathbf{g}(\mathbf{f}(X))]=\mathbb{E}[\mathbf{h}(D)]=\mathbf{0}}} \quad C, \tag{20}$$

where

$$C = L(T(\mathbf{f}(X)), Y) + \lambda \mathbb{E}\Big[\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\, \mathbf{h}(D)\Big] - \frac{\lambda}{2} \operatorname{tr}\big(\mathrm{cov}[\mathbf{g}(\mathbf{f}(X))]\, \mathrm{cov}[\mathbf{h}(D)]\big).$$

We solve this optimization by alternating between optimizing $\mathbf{f}, T$ and optimizing $\mathbf{g}, \mathbf{h}$. In practice, we implement this by alternating between one step of gradient descent for $\mathbf{f}$ and T and five steps of gradient descent on $\mathbf{g}$ and $\mathbf{h}$ to allow the maximal correlation functions to adapt to the changing of features $\mathbf{f}$.

3.2.2. Separation

For separation, we use a similar argument as in the discrete case to ensure the conditional independence. Specifically, we solve the following optimization problem:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} L(T(\mathbf{f}(X)), Y) + \lambda\big(\mathrm{HGR}_{\mathrm{soft}}(f(X), D \otimes Y) - \mathrm{HGR}_{\mathrm{soft}}(f(X), Y)\big). \tag{21}$$

Note that for the first Soft-HGR term, we use $\mathbf{g}, \mathbf{h}$ to denote the maximal correlation functions and $\mathbf{g}', \mathbf{h}'$ to denote the functions for the second term. Similar to the discrete case, the difference term allows us to approximate the conditional mutual information using two unconditional terms. Once again, we solve this optimization by alternating between optimizing $\mathbf{f}$, T and optimizing $\mathbf{g}, \mathbf{h}, \mathbf{g}', \mathbf{h}'$.

3.2.3. Few-Shot Learning

In the continuous case, our learning objective can also be applied a posteriori in a few-shot setting with a clasifier that has already been trained in a fairness-unaware manner on a large number of samples without the sensitive attribute label. In this case, we can formulate our objective as before and use the few samples containing the sensitive attribute to further train the network and force it to learn fairer features that are still predictive of the desired labels.

4. Experimental Results

In order to illustrate the effectiveness of our algorithms, we run experiments using the proposed algorithms on discrete (Adult and COMPAS) and continuous (Communities and Crimes) datasets.

4.1. Discrete Case

We test the proposed DTM-based approach on the ProPublica's COMPAS recidivism dataset (https://github.com/propublica/compas-analysis (accessed on 14 February 2022)) and the UCI Adult dataset (https://archive.ics.uci.edu/ml/datasets/adult (accessed on 14 February 2022)), which were chosen as they contain categorical features and are used in prior works. More experiments for the discrete case can be found in the Appendix A.

For the COMPAS dataset, the goal is to predict whether the individual recidivated (re-offended) (Y) using the severity of charge, number of prior crimes, and age category as the decision variables (X). As discussed in [8], COMPAS scores are biased against African-Americans, so race is set to be the sensitive attribute (D) and filtered to contain only Caucasian and African-American individuals. As for the Adult dataset, the goal is to predict the binary indicator (Y) of whether the income of the individual is more than

50K or not based on the following decision variables (X): age (quantized to decades) and education (in years), and the sensitive attribute (D) is the gender of the individual.

For both datasets, we randomly split all data into 80%/20% training/test samples. We first construct an estimate of DTM $\hat{\mathbf{B}}$ with the empirical distribution of the training set; then, we solve the proposed optimization in (11) and (17) using $\hat{\mathbf{B}}$ to obtain fair feature mappings $\hat{\mathbf{f}}(x), \hat{\mathbf{g}}(y)$. The predictions $\hat{Y}$ of the test samples X' are given by plugging the learned feature mappings $\hat{\mathbf{f}}(x'), \hat{\mathbf{g}}(y)$ into the MAP rule (15), where P_Y can be estimated from the empirical distribution $\hat{P}_Y$ on the training set.

For the independence case, we compare the tradeoff between the performance and the discrimination achieved by our method with that of the optimized pre-processing methods proposed in [8]. Note that we adopt the same settings as the experiments in [8] to do a fair comparison, and the reported results for their method are from their work. We plot the area under the ROC curve (AUC) of $\hat{P}_{Y|X'}(y|x')$ compared to the true test labels Y' against the following standard discrimination measure derived from legal proceedings [4]:

$$J = \max_{d,d' \in \mathcal{D}} \left| \mathbb{P}_{\hat{Y}|D}(1|d) / \mathbb{P}_{\hat{Y}|D}(1|d') - 1 \right|. \tag{22}$$

Figures 1 and 2 (Top) show the results. For both datasets, it can be seen that simply dropping the sensitive attribute D and applying logistic regression (LR) and random forest (RF) algorithms cannot ensure independence between $\hat{Y}$ and D. However, the proposed DTM-based algorithm provides a tradeoff between performance and discrimination by varying the value of the regularizer λ in the optimization (11), which outperforms the optimized pre-processing methods in [8] on the Adult dataset and achieves similar performance on the COMPAS dataset. More importantly, the DTM-based algorithm provides a smooth tradeoff curve between the performance and discrimination, so that a desired level of fairness can be achieved by setting λ in practice. In addition, since our method only requires us to perform eigen-decomposition, it runs significantly faster than the optimized pre-processing method, which needs to solve a much more complex optimization problem. Empirically, we find at least a tenfold speed up in runtime compared to the existing methods.

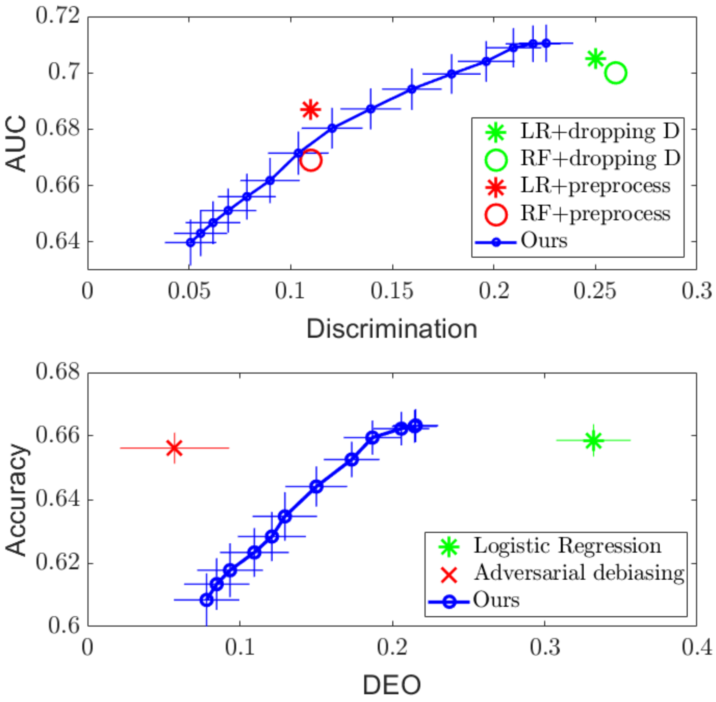

Figure 1. Regularization results on the COMPAS dataset, with AUC plotted against discrimination measure for independence (**Top**), and accuracy plotted against DEO for separation (**Bottom**), respectively.

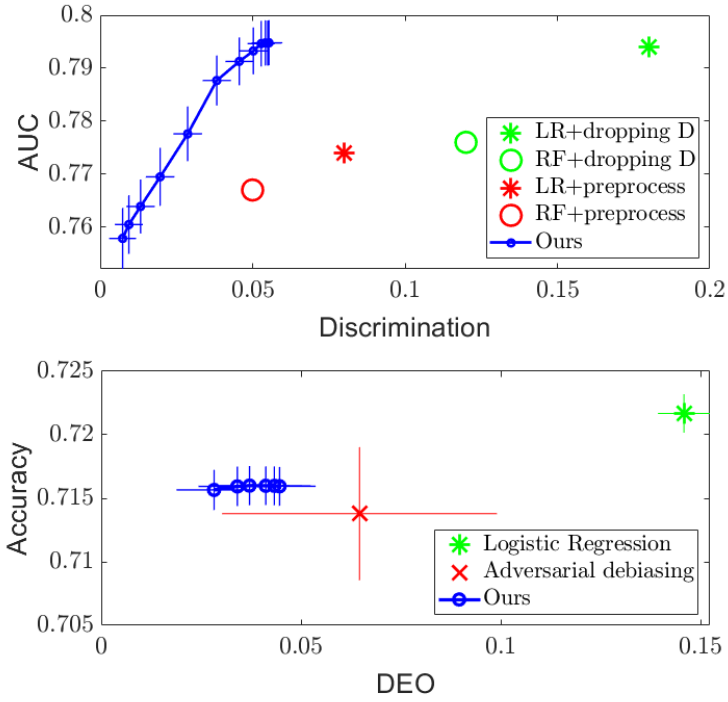

Figure 2. Regularization results on Adult dataset, with AUC plotted against discrimination measure for independence (**Top**), and accuracy plotted against DEO for separation (**Bottom**), respectively.

For the separation criterion, we compare the balanced accuracy achieved by our algorithm with that of the adversarial debiasing method in [20] (implementation given in [2]) against the difference in equalized opportunities (DEO), which is another standard measure used commonly in the literature:

$$\text{DEO} = \left|\mathbb{P}(\hat{Y}{=}1|D{=}1, Y{=}1) - \mathbb{P}(\hat{Y}{=}1|D{=}0, Y{=}1)\right|. \tag{23}$$

The results on the COMPAS and Adult datasets are presented in Figures 1 and 2 (Bottom). Compared to the naïve logistic regression, the proposed DTM-based algorithm dramatically decreases the DEO while maintaining similar accuracy performance on both datasets, which outperforms the adversarial debiasing method in [20] on the Adult dataset. We note that the accuracy and DEO curve achieved by the proposed algorithm in the separation setting has a smaller range compared to that in the independence setting. This is because the value of the regularizer λ is restricted in the separation optimization problem (17) to $\lambda \in [0, 1)$, but only to $\lambda > 0$ for the optimization in (11). More details about the influence of the regularizer λ can be found in Appendix A.

4.2. Continuous Case

In the continuous case, we experiment on the Communities and Crimes (C&C) dataset (http://archive.ics.uci.edu/ml/datasets/communities+and+crime (accessed on 14 February 2022)). The goal is to predict the crime rate Y of a community given a set of 121 statistics X (distributions of income, age, urban/rural, etc.). The 122-th statistic (percentage of black people in the community) is used as the sensitive variable D. All variables in this dataset are real-valued. The dataset was split into 1794 training and 200 test samples. Following [9], we use a Neural Net with a 50-node hidden layer (which we denote as $f(x)$) and train a predictor $\hat{y} = T(f(x))$ with the mean squared error (MSE) loss and the Soft-HGR penalty, varying λ. For Soft-HGR, we use two two-layer NNs with scalar outputs as the two maximal correlation functions $\mathbf{g}$ and $\mathbf{h}$, and then, we trained them according to (20) (independence) or (21) (separation). Then, we computed the test MSE and test “discrimination” in each case.

For independence, our metric was $I(\hat{Y};D)$, which was approximated using a standard kNN-based mutual information estimator [38]. For separation, we computed $I(\hat{Y};D|Y)$ using the same estimator. We report the results of our experiment as well as that of the χ^2 method of [9] with the same architecture. The results of the experiments are presented in Figure 3.

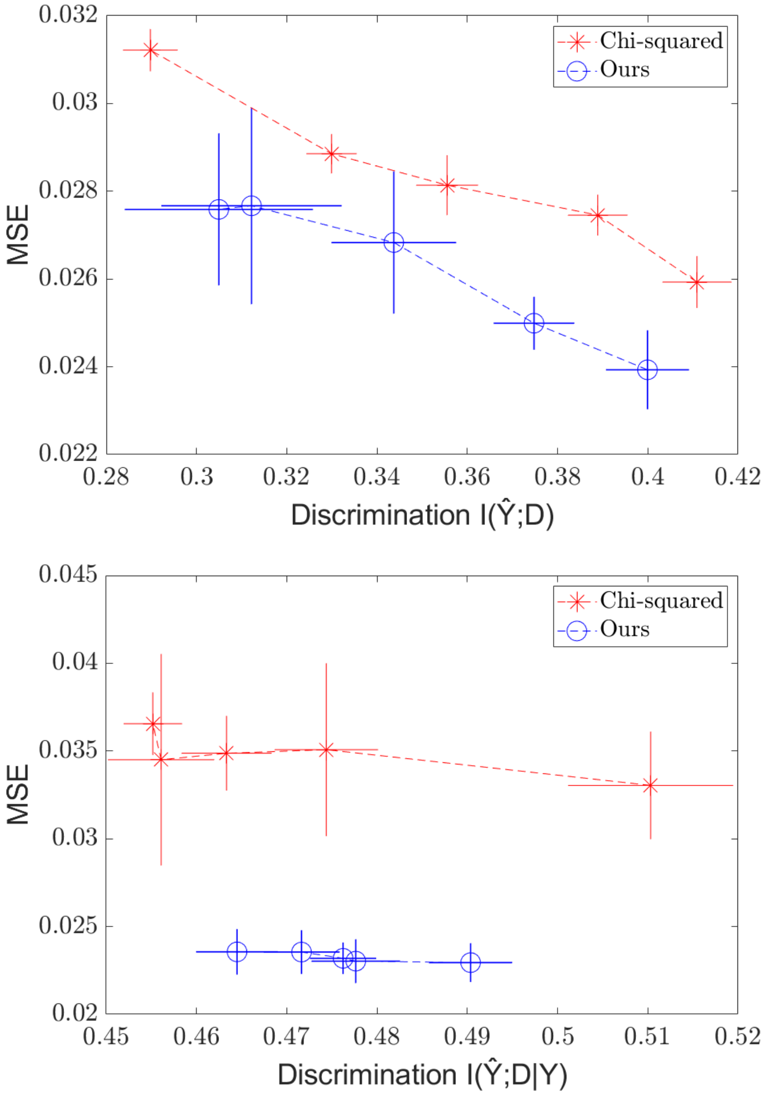

Figure 3. Independence (**top**) and Separation (**bottom**) regularization on the C&C dataset, with MSE plotted against $I(\hat{Y};D|Y)$.

As expected, we see a tradeoff between the MSE and discrimination, creating a frontier of possible values. We also see that the Soft-HGR penalty provides modest gains compared to the χ^2 method for both independence and separation.

Moreover, our method runs significantly faster than the χ^2 method (on the order of seconds per iteration for our method versus just under a minute per iteration for the comparison method), as the χ^2 method requires computation over a mesh grid of a Gaussian KDE, which scales with the product of the number of "bins" (mesh points) and the number of training samples, while our method only scales with the number of samples ($O(n)$), since it only requires passing over all the training samples a constant number of times per iteration. For large bandwidths, d can become quite large. KDE methods also scale poorly with dimensionality (see, [39]) in an exponential manner, and thus, if d is high-dimensional, the χ^2 method would run much slower than our method, which can take in an arbitrarily-sized input and scale linearly with the dimensionality of the input multiplied by the number of samples. Empirically, we find that our method runs around five times faster.

We also run experiments to illustrate how our method's simplicity allows it to adapt to the few-shot, few-epoch regime faster than that of the χ^2 method. We take 10 "few-shot" samples from the training set; then, we train a network to predict Y from X without any fairness regularizer using the full training set. Then, we run five more iterations of gradient descent on the trained model using the fairness-regularized objective and the 10 few-shot samples, and we compare the separation results between the Soft-HGR and χ^2 regularizer. We choose to compare to the χ^2 regularizer as it is one of the few methods designed to handle continuous D. The results are shown in Figure 4. Once again, we see the tradeoff curve, and we see that our method outperform the χ^2 method, and that it appears to be competitive with the standard case in just a few iterations, while the χ^2 method is still far from achieving the original MSE. We also vastly outperform the baseline (before fairness regularization) model in reducing discrimination, at the cost of only a small increase in error. Thus, in situations where, due to ethical/legal issues, only a few samples labeled with the sensitive attribute can be collected, fairness can still be enforced.

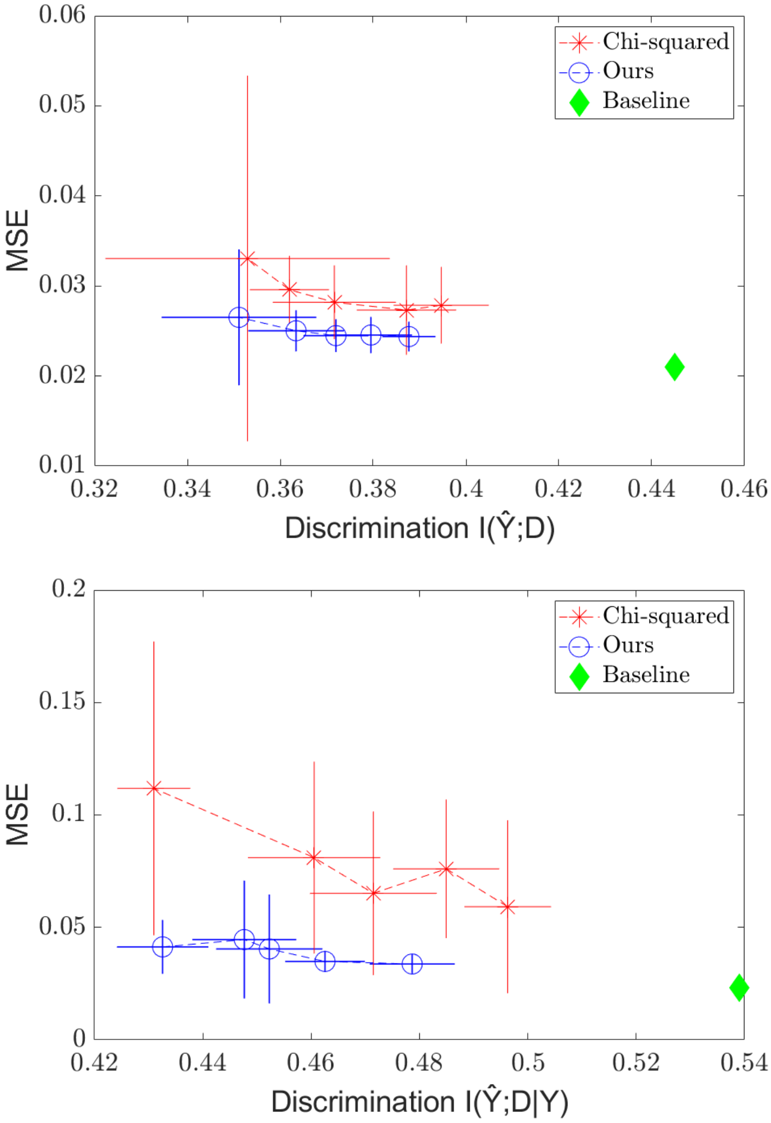

Figure 4. Independence (**top**) and Separation (**bottom**) regularization on the C&C dataset in the *few-shot* settings, with MSE plotted against $I(\hat{Y}; D|Y)$.

5. Conclusions

As machine learning algorithms gain more relevance, more focus will be placed upon ensuring their fairness. We have presented a framework using the HGR maximal correlation, which provides effective and computationally efficient methods for enforcing independence and separation constraints, and derived algorithms for fair learning on discrete and continuous data, which provide competitive tradeoff curves. In addition, we have also shown promising results in the few-shot setting and suggested a method for rapidly adapting a classifier to improve fairness. In the future, it would be beneficial to extend this framework to other criteria (e.g., sufficiency) and to to determine how to use this framework to enforce fairness in a transfer learning setup coupled with the few-shot setting, to determine how to fairly adapt a classifier to a new task.

However, this method requires knowledge of the sensitive attribute for all samples during the training time, which can be impractical in some cases. Further extension into developing these regularizers with a limited number of such samples would be very useful.

Author Contributions: J.L.: software, continuous algorithm design, writing—original draft. Y.B.: software, discrete algorithm design, writing—original draft. P.S. and R.P.: conceptualization, writing—review and editing. G.W.W.: supervision, funding acquisition, and writing—review and editing. L.K. and R.S.F.: writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported, in part, by the MIT-IBM Watson AI Lab under Agreement No. W1771646, and NSF under Grant No. CCF-1717610.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data and code can be found in Section 4.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Effect of the Regularizer in Discrete Case

In this section, we provide additional experiment results to demonstrate how the performance of classification and fairness measures change with different values of the regularizer.

We use the same setup described in Section 4.1 and present the results in Figures A1–A4. In Figures A1 and A2, we plot the achieved AUC and Discrimination (measured with J in (21)) versus the value of λ for both COMPAS and Adult data using independence criterion. In Figures A3 and A4, we plot the accuracy of the classifier and DEO versus λ for both datasets using separation criterion. As shown by all the figures, the performance of classification and fairness measures are all decreasing as we increase λ, and the proposed DTM-based algorithm is able to provide a smooth tradeoff curve between the performance and fairness measures.

Note that the value of the regularizer λ is restricted in the separation optimization problem to $\lambda \in [0, 1)$; therefore, the range of the achieved performance in Figures A3 and A4 is smaller than that in Figures A1 and A2.

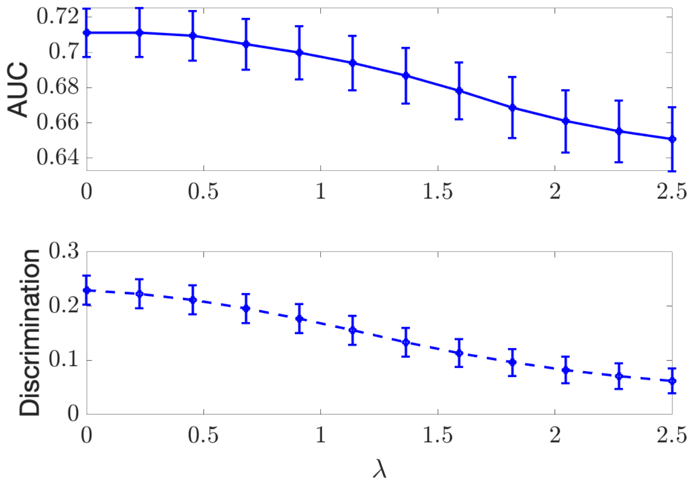

Figure A1. Results for independence regularization on the discrete COMPAS dataset, AUC results (**Top**) and discrimination measure J (**Bottom**) are plotted with respect to different values of λ.

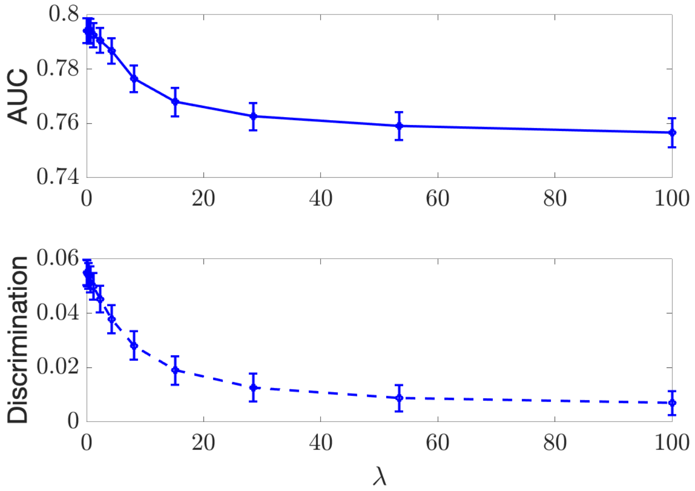

Figure A2. Results for independence regularization on the discrete Adult dataset, AUC results (**Top**) and discrimination measure J (**Bottom**) are plotted with respect to different values of λ.

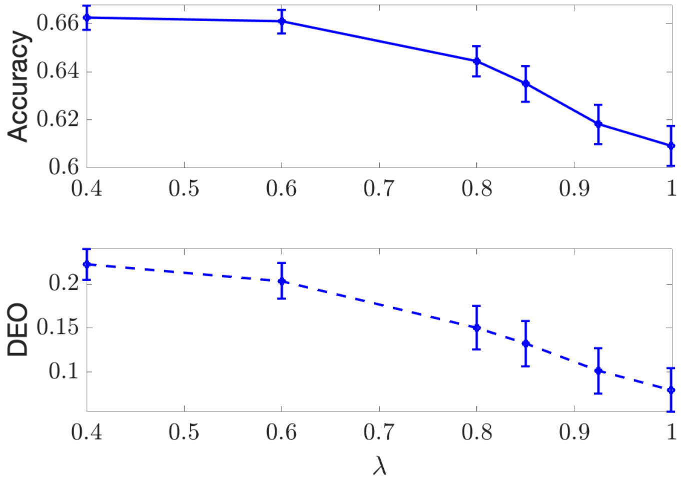

Figure A3. Results for separation regularization on the discrete COMPAS dataset, accuracy (**Top**) and DEO (**Bottom**) are plotted with respect to different values of λ.

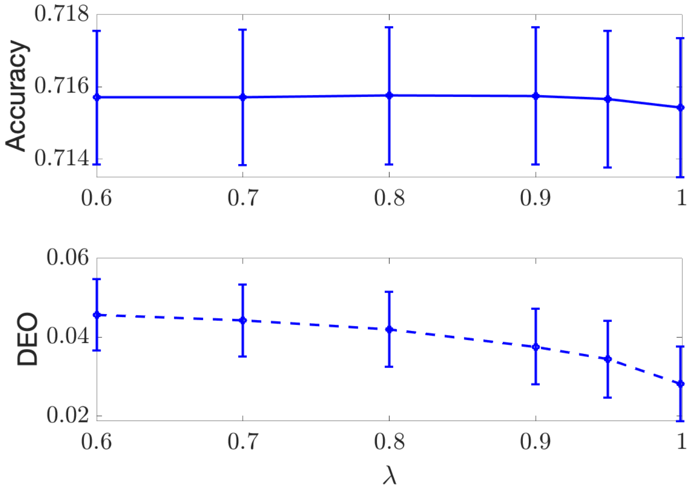

Figure A4. Results for separation regularization on the discrete Adult dataset, accuracy (**Top**) and DEO (**Bottom**) are plotted with respect to different values of λ.

References

1. Selbst, A.D.; Boyd, D.; Friedler, S.A.; Venkatasubramanian, S.; Vertesi, J. Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 59–68.
2. Bellamy, R.K.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv* **2018**, arXiv:1810.01943.
3. Barocas, S.; Hardt, M.; Narayanan, A. Fairness and Machine Learning. 2019. Available online: http://www.fairmlbook.org (accessed on 14 February 2022).
4. EEOC. *Department of Labor, & Department of Justice. Uniform Guidelines on Employee Selection Procedures*; Federal Register: Washington, DC, USA, 1978.
5. Locatello, F.; Abbati, G.; Rainforth, T.; Bauer, S.; Schölkopf, B.; Bachem, O. On the fairness of disentangled representations. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 14584–14597.
6. Gölz, P.; Kahng, A.; Procaccia, A.D. Paradoxes in Fair Machine Learning. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8340–8350.
7. Corbett-Davies, S.; Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* **2018**, arXiv:1808.00023.
8. Calmon, F.; Wei, D.; Vinzamuri, B.; Ramamurthy, K.N.; Varshney, K.R. Optimized pre-processing for discrimination prevention. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 3992–4001.
9. Mary, J.; Calauzenes, C.; El Karoui, N. Fairness-aware learning for continuous attributes and treatments. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 4382–4391.
10. Kamiran, F.; Calders, T. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **2012**, *33*, 1–33. [CrossRef]
11. Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; Dwork, C. Learning fair representations. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 325–333.
12. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and Removing Disparate Impact. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 259–268.
13. Sattigeri, P.; Hoffman, S.C.; Chenthamarakshan, V.; Varshney, K.R. Fairness gan. *arXiv* **2018**, arXiv:1805.09910.
14. Xu, D.; Yuan, S.; Zhang, L.; Wu, X. Fairgan: Fairness-aware generative adversarial networks. In Proceedings of the 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 10–13 December 2018; pp. 570–575.
15. Kamiran, F.; Karim, A.; Zhang, X. Decision Theory for Discrimination-Aware Classification. In Proceedings of the IEEE International Conference on Data Mining, Brussels, Belgium, 10–13 December 2012; pp. 924–929. [CrossRef]
16. Hardt, M.; Price, E.; Srebro, N. Equality of Opportunity in Supervised Learning. In Proceedings of the Advances in Neural Information Processing Systems 29, Barcelona, Spain, 5–10 December 2016; pp. 3315–3323.
17. Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; Weinberger, K.Q. On fairness and calibration. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5680–5689.

18. Wei, D.; Ramamurthy, K.N.; Du Pin Calmon, F. Optimized Score Transformation for Fair Classification. *arXiv* **2019**, arXiv:1906.00066.
19. Kamishima, T.; Akaho, S.; Asoh, H.; Sakuma, J. Fairness-Aware classifier with Prejudice Remover Regularizer. In *Machine Learning and Knowledge Discovery in Databases*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 35–50.
20. Zhang, B.H.; Lemoine, B.; Mitchell, M. Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 335–340.
21. Celis, L.E.; Huang, L.; Keswani, V.; Vishnoi, N.K. Classification with fairness constraints: A meta-algorithm with provable guarantees. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 319–328.
22. Angwin, J.; Larson, J.; Mattu, S.; Kirchner, L. *Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks*; ProPublica: New York, NY, USA, 2016.
23. Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **2017**, *5*, 153–163. [CrossRef] [PubMed]
24. Hirschfeld, H.O. A connection between correlation and contingency. *Proc. Camb. Phil. Soc.* **1935**, *31*, 520–524. [CrossRef]
25. Gebelein, H. Das statistische Problem der Korrelation als Variations-und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Z. Angew. Math. Mech.* **1941**, *21*, 364–379. [CrossRef]
26. Rényi, A. On Measures of Dependence. *Acta Math. Acad. Sci. Hung.* **1959**, *10*, 441–451. [CrossRef]
27. Huang, S.L.; Makur, A.; Wornell, G.W.; Zheng, L. On Universal Features for High-Dimensional Learning and Inference. Preprint, 2019. Available online: http://allegro.mit.edu/~gww/unifeatures (accessed on 14 February 2022).
28. Lee, J.; Sattigeri, P.; Wornell, G. Learning New Tricks From Old Dogs: Multi-Source Transfer Learning From Pre-Trained Networks. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 4372–4382.
29. Wang, L.; Wu, J.; Huang, S.L.; Zheng, L.; Xu, X.; Zhang, L.; Huang, J. An efficient approach to informative feature extraction from multimodal data. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 5281–5288.
30. Rezaei, A.; Fathony, R.; Memarrast, O.; Ziebart, B. Fairness for robust log loss classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5511–5518.
31. Zafar, M.B.; Valera, I.; Rogriguez, M.G.; Gummadi, K.P. Fairness constraints: Mechanisms for fair classification. In Proceedings of theArtificial Intelligence and Statistics, PMLR, Fort Lauderdale, FL, USA, 20–22 April 2017; pp. 962–970.
32. Grari, V.; Ruf, B.; Lamprier, S.; Detyniecki, M. Fairness-Aware Neural Réyni Minimization for Continuous Features. *arXiv* **2019**, arXiv:1911.04929.
33. Baharlouei, S.; Nouiehed, M.; Beirami, A.; Razaviyayn, M. Rènyi Fair Inference. *arXiv* **2019**, arXiv:1906.12005.
34. Moyer, D.; Gao, S.; Brekelmans, R.; Galstyan, A.; Ver Steeg, G. Invariant representations without adversarial training. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 9084–9093.
35. Cho, J.; Hwang, G.; Suh, C. A fair classifier using mutual information. In Proceedings of the 2020 IEEE International Symposium on Information Theory (ISIT), Los Angeles, CA, USA, 21–26 June 2020; pp. 2521–2526.
36. Horn, R.A.; Johnson, C.R. *Matrix Analysis*; Cambridge University Press: Cambridge, UK, 2012.
37. Breiman, L.; Friedman, J.H. Estimating Optimal Transformations for Multiple Regression and Correlation. *J. Am. Stat. Assoc.* **1985**, *80*, 580–598. [CrossRef]
38. Gao, W.; Oh, S.; Viswanath, P. Demystifying Fixed *k*-Nearest Neighbor Information Estimators. *IEEE Trans. Inf. Theory* **2018**, *64*, 5629–5661. [CrossRef]
39. Wang, Z.; Scott, D.W. Nonparametric density estimation for high-dimensional data—Algorithms and applications. *Wiley Interdiscip. Rev. Comput. Stat.* **2019**, *11*, e1461. [CrossRef]