RÉNYI FAIR INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning algorithms have been increasingly deployed in critical automated decision-making systems that directly affect human lives. When these algorithms are solely trained to minimize the training/test error, they could suffer from systematic discrimination against individuals based on their sensitive attributes, such as gender or race. Recently, there has been a surge in machine learning society to develop algorithms for *fair* machine learning. In particular, several adversarial learning procedures have been proposed to impose fairness. Unfortunately, these algorithms either can only impose fairness up to linear dependence between the variables, or they lack computational convergence guarantees. In this paper, we use Rényi correlation as a measure of fairness of machine learning models and develop a general training framework to impose fairness. In particular, we propose a min-max formulation which balances the accuracy and fairness when solved to optimality. For the case of discrete sensitive attributes, we suggest an iterative algorithm with theoretical convergence guarantee for solving the proposed min-max problem. Our algorithm and analysis are then specialized to fair classification and fair clustering problems. To demonstrate the performance of the proposed Rényi fair inference framework in practice, we compare it with well-known existing methods on several benchmark datasets. Experiments indicate that the proposed method has favorable empirical performance against state-of-the-art approaches.

1 INTRODUCTION

As we experience the widespread adoption of machine learning models in automated decision-making, we have witnessed increased reports of instances in which the employed model results in discrimination against certain groups of individuals –see Datta et al. (2015); Sweeney (2013); Bolukbasi et al. (2016); Angwin et al. (2016). In this context, discrimination is defined as the unwanted distinction against individuals based on their membership to a specific group. For instance, Angwin et al. (2016) present an example of a computer-based risk assessment model for recidivism, which is biased against certain ethnicities. In another example, Datta et al. (2015) demonstrate gender discrimination in online advertisements for web pages associated with employment. These observations motivated researchers to pay special attention to fairness in machine learning in recent years; see Calmon et al. (2017); Feldman et al. (2015); Hardt et al. (2016); Zhang et al. (2018); Xu et al. (2018); Dwork et al. (2018); Fish et al. (2016); Woodworth et al. (2017); Zafar et al. (2017; 2015); Pérez-Suay et al. (2017); Bechavod & Ligett (2017).

In addition to its ethical standpoint, equal treatment of different groups is legally required by many countries Act. (1964). Anti-discrimination laws imposed by many countries evaluate fairness by notions such as disparate treatment and disparate impact. We say a decision-making process suffers from disparate treatment if its decisions discriminate against individuals of a certain protected group based on their sensitive/protected attribute information. On the other hand, we say it suffers from disparate impact if the decisions adversely affect a protected group of individuals with certain sensitive attribute – see Zafar et al. (2015). In simpler words, disparate treatment is intentional discrimination against a protected group, while the disparate impact is an unintentional disproportionate outcome that hurts a protected group. To quantify fairness, several notions of fairness have been proposed in the recent decade Calders et al. (2009); Hardt et al. (2016). Examples of these notions include *demographic parity*, *equalized odds*, and *equalized opportunity*.

Demographic parity condition requires that the model output (e.g., assigned label) be independent of sensitive attributes. This definition might not be desirable when the base ground-truth outcome of the two groups are completely different. This shortcoming motivated the use of *equalized odds* notion Hardt et al. (2016) which requires that the model output is conditionally independent of sensitive attributes given the ground-truth label. Finally, *equalized opportunity* requires having equal false positive or false negative rates across protected groups.

Machine learning approaches for imposing fairness can be broadly classified into three main categories: pre-processing methods, post-processing methods, and in-processing methods. Pre-processing methods modify the training data to remove discriminatory information before passing data to the decision-making process Calders et al. (2009); Feldman et al. (2015); Kamiran & Calders (2010; 2009; 2012); Dwork et al. (2012); Calmon et al. (2017); Ruggieri (2014). These methods map the training data to a transformed space in which the dependencies between the class label and the sensitive attributes are removed Edwards & Storkey (2015); Hardt et al. (2016); Xu et al. (2018); Sattigeri et al. (2018); Raff & Sylvester (2018); Madras et al. (2018); Zemel et al. (2013); Louizos et al. (2015). On the other hand, post-processing methods adjust the output of a trained classifier to remove discrimination while maintaining high classification accuracy Fish et al. (2016); Dwork et al. (2018); Woodworth et al. (2017). The third category is the in-process approach that enforces fairness by either introducing constraints or adding a regularization term to the training procedure Zafar et al. (2017; 2015); Pérez-Suay et al. (2017); Bechavod & Ligett (2017); Berk et al. (2017); Agarwal et al. (2018); Celis et al. (2019); Donini et al. (2018); Rezaei et al. (2019); Kamishima et al. (2011); Zhang et al. (2018); Bechavod & Ligett (2017); Kearns et al. (2017); Menon & Williamson (2018); Alabi et al. (2018). The Rényi fair inference framework proposed in this paper also belongs to this in-process category.

Among in-processing methods, many add a regularization term or constraints to promote statistical independence between the classifier output and the sensitive attributes. To do that, various independence proxies such as mutual information Kamishima et al. (2011), false positive/negative rates Bechavod & Ligett (2017), equalized odds Donini et al. (2018), Pearson correlation coefficient Zafar et al. (2015; 2017), Hilbert Schmidt independence criterion Pérez-Suay et al. (2017) were used. As will be discussed in Section 2, many of these methods cannot capture non-linear dependence between random variables and/or lead to computationally expensive algorithms. Motivated by these limitations, we propose to use Rényi correlation to impose several known group fairness measures. Rényi correlation captures nonlinear dependence between random variables. Moreover, Rényi correlation is a normalized measure and can be computed efficiently in certain instances.

Using Rényi correlation coefficient as a regularization term, we propose a min-max optimization framework for fair statistical inference. In particular, We specialize our framework to both classification and clustering tasks. We show that when the sensitive attribute(s) is discrete, the learning task can be efficiently solved to optimality, using a simple gradient ascent-descent approach. We summarize our contributions next:

• We introduce Rényi correlation as a tool to impose several notions of group fairness. Unlike Pearson correlation and HSIC, which only capture linear dependence, Rényi correlation captures any statistical dependence between random variables. Moreover, from computational point of view, it is more efficient than the mutual information regularizers approximated by neural networks.

• Using Rényi correlation as a regularization term in training, we propose a min-max formulation for fair statistical inference. Unlike methods that use an adversarial neural network to impose fairness, we show that in particular instances such as binary classification, or discrete sensitive variable(s), it suffices to use a simple quadratic function as the adversarial objective. This observation helped us to develop a simple multi-step gradient ascent descent algorithm for fair inference and guarantee its theoretical convergence to first-order stationarity.

• Our Rényi correlation framework leads to a natural fair classification method and a novel fair K-means clustering algorithm. For K-means clustering problem, we show that sufficiently large regularization coefficient yields perfect fairness under disparate impact doctrine. Unlike the two-phase methods proposed in Chierichetti et al. (2017); Backurs et al. (2019); Rösner & Schmidt (2018); Bercea et al. (2018); Schmidt et al. (2018), our method does not require any pre-processing step, is scalable, and allows for regulating the trade-off between the clustering quality and fairness.

2 RÉNYI CORRELATION

The most widely used notions for group fairness in machine learning are demographic parity, equalized odds, and equalized opportunities. These notions require (conditional) independence between a certain model output and a sensitive attribute. This independence is typically imposed by adding fairness constraints or regularization terms to the training objective function. For instance, Kamishima et al. (2011) added a regularization term based on mutual information. Since estimating mutual information between the model output and sensitive variables during training is not computationally tractable, Kamishima et al. (2011) approximates the probability density functions using a logistic regression model. To have a tighter estimation, Song et al. (2019) used an adversarial approach that estimates the joint probability density function using a parameterized neural network. Although these works start from a well-justified objective function, they end up solving approximations of the objective function due to computational barriers. Thus, no fairness guarantee can be provided even when the resulting optimization problems are solved to global optimality in the large sample size scenarios. A more tractable measure of dependence between two random variables is the Pearson correlation. The Pearson correlation coefficient between the two random variables A and B is defined as $\rho_P(A,B) = \frac{\mathrm{Cov}(A,B)}{\sqrt{\mathrm{Var}(A)}\sqrt{\mathrm{Var}(B)}}$, where $\mathrm{Cov}(\cdot,\cdot)$ denotes the covariance and $\mathrm{Var}(\cdot)$ denotes the variance. The Pearson correlation coefficient is used in Zafar et al. (2015) to decorrelate the binary sensitive attribute and the decision boundary of the classifier. A major drawback of Pearson correlation is that it only captures linear dependencies between random variables. In fact, two random variables A and B may have strong dependence but have zero Pearson correlation. This property raises concerns about the use of the Pearson correlation for imposing fairness. Similar to the Pearson correlation, the HSIC measure proposed in Pérez-Suay et al. (2017) may be zero even if the two variables have strong dependencies. While universal Kernels can be used to resolve this issue, they could arrive at the expense of computational interactability. In addition, HSIC is not a normalized dependence measure Gretton et al. (2005b;a) which raises concerns about the appropriateness of using it as a measure of dependence.

In this paper, we suggest to use Hirschfeld-Gebelein-Rényi correlation Rényi (1959); Hirschfeld (1935); Gebelein (1941) as a dependence measure between random variables to impose fairness. Rényi correlation, which is also known as maximal correlation, between two random variables A and B is defined as

$$\begin{aligned}\rho_R(A,B) = &\sup_{f,g} \mathbb{E}[f(A)g(B)] \\ &\text{s.t.} \quad \mathbb{E}[f(A)] = \mathbb{E}[g(B)] = 0, \quad \mathbb{E}[f^2(A)] = \mathbb{E}[g^2(B)] = 1,\end{aligned} \tag{1}$$

where the supremum is over the set of measurable functions $f(\cdot)$ and $g(\cdot)$ satisfying the constraints. Unlike HSIC and Pearson correlation, Rényi correlation is a normalized measure that captures higher-order dependencies between random variables. Rényi correlation between two random variables is zero if and only if the random variables are independent, and it is one if there is a strict dependence between the variables Rényi (1959). These favorable statistical properties of ρ_R do not come at the price of computational intractability. In fact, as we will discuss in Section 3, ρ_R can be used in a computationally tractable framework to impose several group fairness notions.

3 A GENERAL MIN-MAX FRAMEWORK FOR RÉNYI FAIR INFERENCE

Consider a learning task over a given random variable $\mathbf{Z}$. Our goal is to minimize the average inference loss $\mathcal{L}(\cdot)$ where our loss function is parameterized with parameter $\boldsymbol{\theta}$. To find the optimal value of parameter $\boldsymbol{\theta}$ with the smallest average loss, we need to solve the following optimization problem

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}\big[\mathcal{L}(\boldsymbol{\theta}, \mathbf{Z})\big],$$

where the expectation is taken over $\mathbf{Z}$. Notice that this formulation is quite general and can include regression, classification, clustering, or dimensionality reduction tasks as special cases. As an example, in the case of linear regression $\mathbf{Z} = (\mathbf{X}, Y)$ and $\mathcal{L}(\boldsymbol{\theta}, \mathbf{Z}) = (Y - \boldsymbol{\theta}^T\mathbf{X})^2$ where $\mathbf{X}$ is a random vector and $\mathbf{Y}$ is the random target variable.

Assume that, in addition to minimizing the average loss, we are interested in bringing fairness to our learning task. Let S be the sensitive attribute and $\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z})$ be a certain output of our inference task using parameter $\boldsymbol{\theta}$. Assume we are interested in reducing the dependence between the random variable $\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z})$ and the sensitive attribute S. To balance the goodness-of-fit and fairness, one can solve the following optimization problem

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}\big[\mathcal{L}(\boldsymbol{\theta}, \mathbf{Z})\big] + \lambda\, \rho_R^2\big(\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z}), S\big), \tag{2}$$

where λ is a positive scalar balancing fairness and goodness-of-fit. Notice that the above framework is quite general. For example, $\hat{Y}_\theta$ may be the assigned label in a classification task, the assigned cluster in a clustering task, or the output of a regressor in a regression task.

Using the definition of Rényi correlation, we can rewrite optimization problem equation 2 as

$$\begin{aligned} \min_{\boldsymbol{\theta}} \sup_{f,g} \quad & \mathbb{E}\big[\mathcal{L}(\boldsymbol{\theta}, \mathbf{Z})\big] + \lambda \left(\mathbb{E}\big[f(\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z}))\, g(S)\big]\right)^2, \\ \text{s.t.} \quad & \mathbb{E}\big[f(\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z}))\big] = \mathbb{E}\big[g(S)\big] = 0, \quad \mathbb{E}\big[f^2(\widehat{Y}_{\boldsymbol{\theta}}(\mathbf{Z}))\big] = \mathbb{E}\big[g^2(S)\big] = 1, \end{aligned} \tag{3}$$

where the supremum is taken over the set of measurable functions. The next natural question to ask is whether this optimization problem can be efficiently solved in practice. This question motivates the discussions of the following subsection.

3.1 Computing Rényi Correlation

The objective function in equation 3 may be non-convex in $\boldsymbol{\theta}$ in general. Several algorithms have been recently proposed for solving such non-convex min-max optimization problems Sanjabi et al. (2018); Nouiehed et al. (2019); Jin et al. (2019). Most of these methods require solving the inner maximization problem to (approximate) global optimality. More precisely, we need to be able to solve the optimization problem described in equation 1. While popular heuristic approaches such as parameterizing the functions f and g with neural networks can be used to solve equation 1, we focus on solving this problem in a more rigorous manner. In particular, we narrow down our focus to the discrete random variable case. This case holds for many practical sensitive attributes among which are the gender and race. In what follows, we show that in this case, equation 1 can be solved "efficiently" to global optimality.

Theorem 3.1 (Witsenhausen (1975)). *Let $a \in \{a_1, \ldots, a_c\}$ and $b \in \{b_1, \ldots, b_d\}$ be two discrete random variables. Then the Rényi coefficient $\rho_R(a, b)$ is equal to the second largest singular value of the matrix $\mathbf{Q} = [q_{ij}]_{i,j} \in \mathbb{R}^{c \times d}$, where $q_{ij} = \frac{\mathbb{P}(a=a_i, b=b_j)}{\sqrt{\mathbb{P}(a=a_i)\mathbb{P}(b=b_j)}}$.*

The above theorem provides a computationally tractable approach for computing the Rényi coefficient. This computation could be further simplified when one of the random variables is binary.

Theorem 3.2. *Suppose that $a \in \{1, \ldots, c\}$ is a discrete random variable and $b \in \{0, 1\}$ is a binary random variable. Let $\tilde{\mathbf{a}}$ be the one-hot encoded version of a, i.e., $\tilde{\mathbf{a}} = \mathbf{e}_i$ if $a = i$, where $\mathbf{e}_i = (0, \ldots, 0, 1, 0 \ldots, 0)$ is the i-th standard unit vector. Let $\tilde{b} = b - 1/2$. Then,*

$$\rho_R^2(a, b) \triangleq 1 - \frac{\gamma}{\mathbb{P}(b=1)\mathbb{P}(b=0)}, \tag{4}$$

where $\gamma \triangleq \min_{\mathbf{w} \in \mathbb{R}^c} \quad \mathbb{E}\left[(\mathbf{w}^T \widetilde{\mathbf{a}} - \tilde{b})^2\right]$. Equivalently,

$$\gamma \triangleq \min_{\mathbf{w} \in \mathbb{R}^c} \quad \sum_{i=1}^{c} w_i^2 \mathbb{P}(a=i) - \sum_{i=1}^{c} w_i \big(\mathbb{P}(a=i, b=1) - \mathbb{P}(a=i, b=0)\big) + 1/4. \tag{5}$$

Proof. The proof is relegated to the appendix section. □

Let us specialize our framework to classification and clustering problems in the next two sections.

4 RÉNYI FAIR CLASSIFICATION

In a typical (multi-class) classification problem, we are given samples from a random variable $\mathbf{Z} \triangleq (\mathbf{X}, Y)$ and the goal is to predict Y from $\mathbf{X}$. Here $\mathbf{X} \in \mathbb{R}^d$ is the input feature vector, and $Y \in \mathcal{Y} \triangleq \{1, \ldots, c\}$ is the class label. Let $\widehat{Y}_{\boldsymbol{\theta}} \in \{1, \ldots, c\}$ be the output of our classifier with

$$\mathbb{P}(\widehat{Y}_{\boldsymbol{\theta}} = i \mid \mathbf{X}) = \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{X}), \quad \forall i = 1, \ldots, c.$$

Here $\boldsymbol{\theta}$ is that parameter of the classifier that needs to be tuned. For example, $\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}) = (\mathcal{F}_1(\boldsymbol{\theta}, \mathbf{X}), \ldots, \mathcal{F}_c(\boldsymbol{\theta}, \mathbf{X}))$ could represent the output of a neural network after *softmax* layer; or the soft probability label assigned by a logistic regression model. In order to find the optimal parameter $\boldsymbol{\theta}$, we need to solve the optimization problem

$$\min_{\boldsymbol{\theta}} \ \mathbb{E}\Big[\mathcal{L}(\mathcal{F}\big(\boldsymbol{\theta}, \mathbf{X}\big), Y)\Big], \tag{6}$$

where $\mathcal{L}$ is the loss function and the expectation is taken over the random variable $\mathbf{Z} = (\mathbf{X}, Y)$. Let S be the sensitive attribute. We say a model satisfies *demographic parity* if the assigned label $\widehat{Y}$ is independent of the sensitive attribute S, see Dwork et al. (2012). Using our regularization framework, to find the optimal parameter $\boldsymbol{\theta}$ balancing classification accuracy and fairness objective, we need to solve

$$\min_{\boldsymbol{\theta}} \ \mathbb{E}\Big[\mathcal{L}\big(\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}), Y\big)\Big] + \lambda \rho_R^2\big(\widehat{Y}_{\boldsymbol{\theta}}, S\big). \tag{7}$$

4.1 GENERAL DISCRETE CASE

When $S \in \{s_1, \ldots, s_d\}$ is discrete, Theorem 3.1 implies that equation 7 can be rewritten as

$$\min_{\boldsymbol{\theta}} \ \max_{\mathbf{v} \perp \mathbf{v}_1, \, \|\mathbf{v}\|^2 \leq 1} \left(f_D(\boldsymbol{\theta}, \mathbf{v}) \triangleq \mathbb{E}\Big[\mathcal{L}\big(\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}), \mathbf{Y}\big)\Big] + \lambda \mathbf{v}^T \mathbf{Q}_{\boldsymbol{\theta}}^T \mathbf{Q}_{\boldsymbol{\theta}} \mathbf{v} \right). \tag{8}$$

Here $\mathbf{v}_1 = \Big[\sqrt{\mathbb{P}(S = s_1)}, \ldots, \sqrt{\mathbb{P}(S = s_d)}\Big] \in \mathbb{R}^d$ is the right singular vector corresponding to the largest singular value of $\mathbf{Q}_{\boldsymbol{\theta}} = [q_{ij}]_{i,j} \in \mathbb{R}^{c \times d}$, with $q_{ij} \triangleq \dfrac{\mathbb{P}(\widehat{Y}_{\boldsymbol{\theta}} = i \,|\, S = s_j)\mathbb{P}(S = s_j)}{\sqrt{\mathbb{P}(\widehat{Y}_{\boldsymbol{\theta}} = i) \, \mathbb{P}(S = s_j)}}$.

Given training data $(\mathbf{x}_n, y_n)_{n=1}^N$ sampled from the random variable $\mathbf{Z} = (\mathbf{X}, Y)$, we can estimate the entries of the matrix $\mathbf{Q}_{\boldsymbol{\theta}}$ using $\mathbb{P}\big(\widehat{Y}_{\boldsymbol{\theta}} = i\big) = \mathbb{E}[\mathbb{P}(\widehat{Y}_{\boldsymbol{\theta}} = i \mid \mathbf{X})] \approx \dfrac{1}{N} \sum_{n=1}^N \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}_n)$, and $\mathbb{P}\big(\widehat{Y}_{\boldsymbol{\theta}} = i \,| S = s_j\big) \approx \dfrac{1}{|\mathcal{X}_j|} \sum_{\mathbf{x} \in \mathcal{X}_j} \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x})$, where $\mathcal{X}_j$ is the set of samples with sensitive attribute s_j. Motivated by the algorithm proposed in Jin et al. (2019), we present Algorithm 1 for solving equation 8.

Algorithm 1 Rényi Fair Classifier for Discrete Sensitive Attributes

1: **Input**: $\boldsymbol{\theta}^0 \in \Theta$, step-size η.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: Set $\mathbf{v}^{t+1} \leftarrow \max_{\mathbf{v} \in \perp \mathbf{v}_1, \|\mathbf{v}\| \leq 1} f_D(\boldsymbol{\theta}^t, \mathbf{v})$ by finding the second singular vector of $\mathbf{Q}_{\boldsymbol{\theta}^t}$
4: Set $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f_D(\boldsymbol{\theta}^t, \mathbf{v}^{t+1})$
5: **end for**

To understand the convergence behavior of Algorithm 1 for the nonconvex optimization problem equation 8, we need to first define an approximate stationary solution. Let us define $g(\boldsymbol{\theta}) = \max_{\mathbf{v} \in \perp \mathbf{v}_1, \|\mathbf{v}\| \leq 1} f(\boldsymbol{\theta}, \mathbf{v})$. Assume further that $f(\cdot, \mathbf{v})$ has L_1-Lipschitz gradient, then $g(\cdot)$ is L_1-weakly convex; for more details check Rafique et al. (2018). For such weakly convex function, we say $\boldsymbol{\theta}^*$ is a ϵ-stationary solution if the gradient of its Moreau envelop is smaller than epsilon, i.e., $\|\nabla g_\beta(\cdot)\| \leq \varepsilon$ with $g_\beta(\boldsymbol{\theta}) \triangleq \min_{\boldsymbol{\theta}'} g(\boldsymbol{\theta}') + \dfrac{1}{2\beta}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ and $\beta < \dfrac{1}{2L_1}$ is a given constant. The following theorem demonstrate the convergence of Algorithm 1. This theorem is the direct consequence of Theorem 27 in Jin et al. (2019).

Theorem 4.1. *Suppose that f is L_0-Lipschitz and L_1-gradient Lipschitz. Then Algorithm 1 computes an ε-stationary solution of the objective function in equation 8 in $\mathcal{O}(\varepsilon^{-4})$ iterations.*

4.2 Binary Case

When S is binary, we can obtain a more efficient algorithm compared to Algorithm 1 by exploiting Theorem 3.2. Particularly, by a simple scaling of λ and ignoring the constant terms, the optimization problem equation 7 can be written as

$$\begin{aligned}\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \quad f(\boldsymbol{\theta}, \mathbf{v}) \triangleq \;& \mathbb{E}\big[\mathcal{L}\big(\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}), Y\big)\big] - \lambda\Big[\textstyle\sum_{i=1}^{c} w_i^2 \mathbb{P}\Big(\widehat{Y}_{\boldsymbol{\theta}} = i\Big) \\ & - \textstyle\sum_{i=1}^{c} w_i \Big(\mathbb{P}\big(\widehat{Y}_{\boldsymbol{\theta}} = i, S = 1\big) - \mathbb{P}\big(\widehat{Y}_{\boldsymbol{\theta}} = i, S = 0\big)\Big)\Big].\end{aligned} \tag{9}$$

Defining $\tilde{S} = 2S - 1$, the above problem can be rewritten as

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \quad \mathbb{E}\big[\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}), Y) - \lambda \sum_{i=1}^{c} w_i^2 \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{X}) + \lambda \sum_{i=1}^{c} w_i \widetilde{S} \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{X})\big]$$

Thus, given training data $(\mathbf{x}_n, y_n)_{n=1}^{N}$ sampled from the random variable $\mathbf{Z} = (\mathbf{X}, Y)$, we solve

$$\min_{\boldsymbol{\theta}} \max_{\mathbf{w}} \left[f_B(\boldsymbol{\theta}, \mathbf{w}) \triangleq \frac{1}{N} \sum_{n=1}^{N} \big[\mathcal{L}\big(\mathcal{F}(\boldsymbol{\theta}, \mathbf{x}_n), y_n\big) - \lambda \sum_{i=1}^{c} w_i^2 \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}_n) + \lambda \sum_{i=1}^{c} w_i \widetilde{s}_n \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}_n)\big]\right] \tag{10}$$

Notice that the maximization problem in equation 10 is concave, separable, and has a closed-form solution. We propose Algorithm 2 for solving equation 10.

Algorithm 2 Rényi Fair Classifier for Binary Sensitive Attributes

1: **Input**: $\boldsymbol{\theta}^0 \in \Theta$, step-size η.
2: **for** $t = 0, 1, \ldots, T$ **do**
3: Set $\mathbf{w}^{t+1} \leftarrow \arg\max_{\mathbf{w}} f_B(\boldsymbol{\theta}^t, \mathbf{w})$, i.e., set $w_i^{t+1} \leftarrow \frac{\sum_{n=1}^{N} \widetilde{s}_n \mathcal{F}_i(\boldsymbol{\theta}^t, \mathbf{x}_n)}{2\sum_{n=1}^{N} \mathcal{F}_i(\boldsymbol{\theta}^t, \mathbf{x}_n)}, \quad \forall i = 1, \ldots, c$
4: Set $\boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla_{\boldsymbol{\theta}} f_B(\boldsymbol{\theta}^t, \mathbf{w}^{t+1})$
5: **end for**

While the result in Theorem 4.1 can be applied to Algorithm 2, under the following assumption, we can show a superior convergence rate.

Assumption 4.1. *We assume that there exists a constant scalar $\mu > 0$ such that $\sum_{n=1}^{N} \mathcal{F}_i(\boldsymbol{\theta}, \mathbf{x}_n) \geq \mu, \quad \forall i = 1, \ldots, C$.*

This assumption is reasonable when soft-max is used. This is because we can always assume $\boldsymbol{\theta}$ lies in a compact set in practice, and hence the output of the softmax layer cannot be arbitrarily small.

Theorem 4.2. *Suppose that f is L_1-gradient Lipschitz. Then Algorithm 2 computes an ε-stationary solution of the objective function in equation 10 in $\mathcal{O}(\varepsilon^{-2})$ iterations.*

Proof. The proof is relegated to the appendix section. □

Notice that this convergence rate is clearly a faster rate than the one obtained in Theorem 4.1.

Remark 4.3 (Extension to multiple sensitive attributes)**.** *Our discrete Rényi classification framework can naturally be extended to the case of multiple discrete sensitivity attributes by concatenating all attributes into one. For instance, when we have two sensitivity attribute $S^1 \in \{0, 1\}$ and $S^2 \in \{0, 1\}$, we can consider them as a single attribute $S \in \{0, 1, 2, 3\}$ corresponding to the four combinations of $\{(S^1 = 0, S^2 = 0), (S^1 = 0, S^2 = 1), (S^1 = 0, S^2 = 0), (S^1 = 1, S^2 = 1)\}$.*

Remark 4.4 (Extension to other notions of fairness)**.** *Our proposed framework imposes the demographic parity notion of group fairness. However, other notions of group fairness may be represented by (conditional) independence conditions. For such cases, we can again apply our framework. For example, we say a predictor $\hat{Y}_{\boldsymbol{\theta}}$ satisfies equalized odds condition if the predictor $\hat{Y}_{\boldsymbol{\theta}}$ is conditionally* *independent of the sensitive attribute S given the true label Y. Similar to formulation equation 7, the equalized odds fairness notion can be achieved by the following min-max problem*

$$\min_{\boldsymbol{\theta}} \quad \mathbb{E}\Big[\mathcal{L}\big(\mathcal{F}(\boldsymbol{\theta}, \mathbf{X}), Y\big)\Big] + \lambda \sum_{y \in \mathcal{Y}} \rho_R^2 \left(\hat{Y}_{\boldsymbol{\theta}}, S \mid Y = y\right). \tag{11}$$

5 Rényi Fair Clustering

In this section, we apply the proposed fair Rényi framework to the widespread K-means clustering problem. Given a set of data points $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^{N \times d}$, in the K-means problem, we seek to partition them into K clusters such that the following objective function is minimized:

$$\min_{\mathbf{A}, \mathbf{C}} \sum_{n=1}^{N} \sum_{k=1}^{K} a_{kn} \|\mathbf{x}_n - \mathbf{c}_k\|^2 \quad \text{s.t.} \quad \sum_{k=1}^{K} a_{kn} = 1, \ \forall n, \quad a_{kn} \in \{0, 1\}, \ \forall k, n \tag{12}$$

where $\mathbf{c}_k$ is the centroid of cluster k; the variable $a_{kn} = 1$ if data point $\mathbf{x}_n$ belongs to cluster k and it is zero otherwise; $\mathbf{A} = [a_{kn}]_{k,n}$ and $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_K]$ represent the association matrix and the cluster centroids respectively. Now, suppose we have an additional sensitive attribute S for each one of the given data points. In order to have a fair clustering under disparate impact doctrine, we need to make the random variable $\mathbf{a}_n = [a_{1n}, \ldots, a_{Kn}]$ independent of S. In other words, we need to make the clustering assignment independent of the sensitive attribute S. Using our framework in equation 2, we can easily add a regularizer to this problem to impose fairness under disparate impact doctrine. In particular, for binary sensitive attribute S, using Theorem 3.2, and absorbing the constants into the hyper-parameter λ, we need to solve

$$\begin{aligned} &\min_{\mathbf{A}, \mathbf{C}} \max_{\mathbf{w} \in \mathbb{R}^K} \sum_{n=1}^{N} \sum_{k=1}^{K} a_{kn} \|\mathbf{x}_n - \mathbf{c}_k\|^2 - \lambda \sum_{n=1}^{N} (\mathbf{a}_n^T \mathbf{w} - s_n)^2 \\ &\text{s.t.} \qquad \textstyle\sum_{k=1}^{K} a_{kn} = 1, \quad \forall\, n, \quad a_{kn} \in \{0, 1\}, \quad \forall\, k, n. \end{aligned} \tag{13}$$

where $\mathbf{a}_n = (a_{1n}, \ldots, a_{Kn})^T$ encodes the clustering information of data point $\mathbf{x}_n$ and s_n is the sensitive attribute for data point n.

Fixing the assignment matrix $\mathbf{A}$, and cluster centers $\mathbf{C}$, the vector $\mathbf{w}$ can be updated in closed-form. More specifically, w_k at each iteration equals to the current proportion of the privileged group in the k-th cluster. Combining this idea with the update rules of assignments and cluster centers in the standard K-means algorithm, we propose Algorithm 3, which is a fair K-means algorithm under disparate impact doctrine. To illustrate the behavior of the algorithm, a toy example is presented in Appendix C.

Algorithm 3 Rényi Fair K-means

1: **Input**: $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{S} = \{s_1, \ldots, s_N\}$
2: **Initialize**: Random assignment $\mathbf{A}$ s.t. $\sum_{k=1}^{K} a_{kn} = 1 \, \forall\, n;$ and $a_{kn} \in \{0, 1\}$. $Set\ \mathbf{A}_{prev} = \mathbf{0}$.
3: **while** $\mathbf{A}_{prev} \neq \mathbf{A}$ **do**
4: Set $\mathbf{A}_{prev} = \mathbf{A}$
5: **for** $n = 1, \ldots, N$ **do** ▷ Update $\mathbf{A}$
6: $k^* = \arg\min_k \|\mathbf{x}_n - \mathbf{c}_k\|^2 - \lambda(\mathbf{w}_k - s_n)^2$
7: Set $a_{k^* n} = 1$ and $a_{kn} = 0$ for all $k \neq k^*$
8: Set $w_k = \dfrac{\sum_{n=1}^{N} s_n a_{kn}}{\sum_{n=1}^{N} a_{kn}}, \ \forall k = 1, \ldots, K.$ ▷ Update $\mathbf{w}$
9: **end for**
10: Set $\mathbf{c}_k = \dfrac{\sum_{n=1}^{N} a_{kn} \mathbf{x}_n}{\sum_{n=1}^{N} a_{kn}}, \ \forall k = 1, \ldots, K.$ ▷ Update $\mathbf{c}$
11: **end while**

The main difference between this algorithm and the popular K-means algorithm is in Step 6 of Algorithm 3. This step is a result of optimizing equation 13 over $\mathbf{A}$ when both $\mathbf{C}$ and $\mathbf{w}$ are fixed.

When $\lambda = 0$, this step would be identical to the update of cluster assignment variables in K-means. However, when $\lambda > 0$, Step 6 considers fairness when computing the distance considered in updating the cluster assignments.

Remark 5.1. *Note that in Algorithm 3, the parameter* $\mathbf{w}$ *is being updated after each assignment of a point to a cluster. More specifically, for every iteration of the algorithm,* $\mathbf{w}$ *is updated* N *times. If we otherwise update* $\mathbf{w}$ *after completely updating the matrix* $\mathbf{A}$*, then with a simple counterexample we can show that the algorithm can get stuck; see more details in Appendix C.1.*

6 Numerical Experiments

In this section, we evaluate the performance of the proposed Rényi fair classifier and Rényi fair k-means algorithm on three standard datasets: *Bank*, *German Credit*, and *Adult* datasets. The detailed description of these datasets is available in the supplementary material.

We evaluate the performance of our proposed Rényi classifier under both demographic parity, and equality of opportunity notions. We have implemented a logistic regression classifier regularized by Rényi correlation on Adult dataset considering gender as the sensitive feature. To measure the equality of opportunity we use the Equality of Opportunity (EO) violation, defined as $\text{EO Violation} = \big|\mathbb{P}(\hat{Y} = 1|S = 1, Y = 1) - \mathbb{P}(\hat{Y} = 1|S = 0, Y = 1)\big|$, where $\hat{Y}$ and Y represent the predicted, and true labels respectively. Smaller EO violation corresponds to a more fair solution. Figure 1, part (a) and (b) demonstrate that by increasing λ, the Rényi regularizer coefficient decreases implying a more fair classifier at the price of a higher training and testing errors. Figure 1, part (c) compares the fair Rényi logistic regression model to several existing methods in the literature Hardt et al. (2016); Zafar et al. (2015); Rezaei et al. (2019); Donini et al. (2018). As we can see in plot (c), Rényi classifier outperforms other methods in terms of accuracy for a given level of EO violation.

To show the practical benefits of Rényi over Pearson and HSIC regularizers under the demographic parity notion, we evaluate the logistic regression classifier regularized by these three measures on Adult, Bank, and German Credit datasets. For the first two plots, we use $p\% = \min(\frac{\mathbb{P}(\hat{Y}=1|S=1)}{\mathbb{P}(\hat{Y}=1|S=0)}, \frac{\mathbb{P}(\hat{Y}=1|S=0)}{\mathbb{P}(\hat{Y}=1|S=1)})$ as a measure of fairness. Since $p\%$ is defined only for binary sensitive variables, for the last two plots in Figure 2 (German dataset with gender and marital status, and Adult dataset with gender and race as the sensitive features), we use the inverse of demographic parity (DP) violation as the fairness measure. We define DP violation as $\text{DP Violation} = \max_{a,b}\big|\mathbb{P}(\hat{Y} = 1|S = a) - \mathbb{P}(\hat{Y} = 1|S = b)\big|$. As it is evident from the figure, Rényi classifier outperforms both HSIC and Pearson classifiers, especially when targeting high levels of fairness. For the last two experiments in Figure 2, we could not further increase fairness by increasing the regularization coefficient for Pearson and HSIC regularizers, see green and red curves cannot go beyond a certain point on the fairness axis. This can be explained by the non-linear correlation between the predictor and the sensitive variables in these two scenarios which cannot be fully captured using linear or quadratic independence measures. Interestingly, our experiments indicate that minimizing Rényi correlation eventually minimizes the Normalized Mutual Information (NMI) between the variables (See Supplementary Figure6). Recall that similar to Rényi correlation, NMI can capture any dependence between two given random variables.

Finally, to evaluate the performance of our fair k-means algorithm, we implement Algorithm 3 to find clusters of Adult and Bank datasets. We use the deviation of the elements of the vector $\mathbf{w}$ as a measure of fairness. The element $\mathbf{w}_k$ of $\mathbf{w}$ represents the ratio of the number of data points that belong to the privileged group ($S = 1$) in cluster k over the number of data points in that cluster. This notion of fairness is closely related to *minimum balance* introduced by Chierichetti et al. (2017). The deviation of these elements is a measure for the deviation of these ratios across different clusters. A clustering solution is exactly fair if all entries of $\mathbf{w}$ are the same. For $K = 14$, we plot in Figure 3 the minimum, maximum, average, and average $\pm$ standard deviation of the entries of $\mathbf{w}$ vector for different values of λ. For an exactly fair clustering solution, these values should be the same. As we can see in Figure 3, increasing λ yields exact fair clustering at the price of a higher clustering loss.

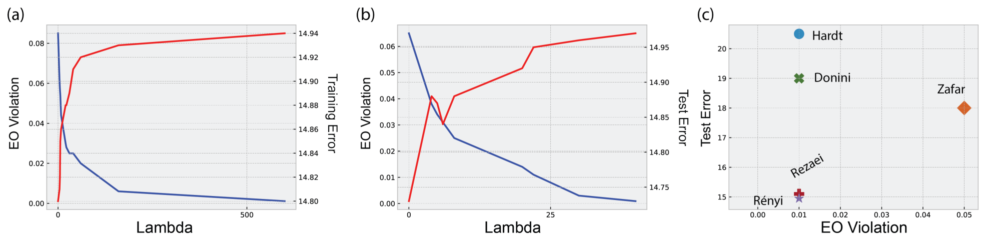


Figure 1: Trade-off between the accuracy of classifier and fairness on the adult dataset under the equality of opportunity notion.

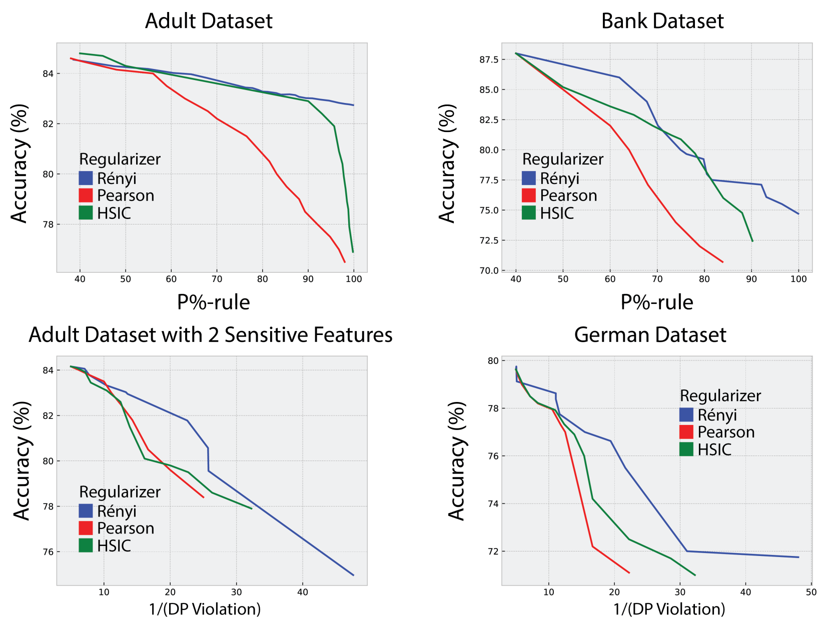


Figure 2: Trade-off between accuracy and fairness for logistic regression classifier regularized with Rényi , HSIC, and Pearson measures, on German Credit, Adult, and Bank datasets.

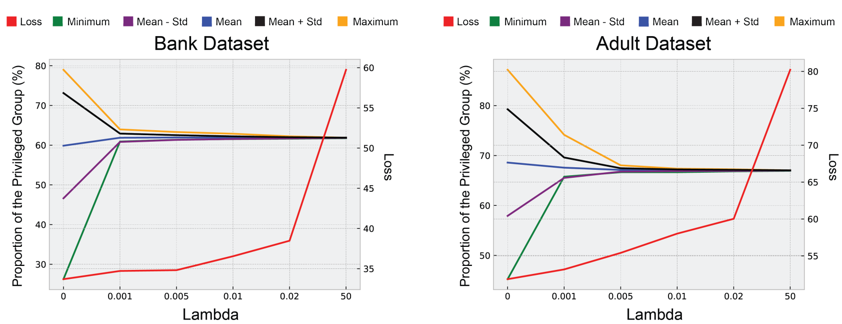


Figure 3: Performance and fairness of K-means algorithm in terms of Rényi regularizer hyper-parameter λ.

References

Civil Rights Act. Civil rights act of 1964, title vii,equal employment opportunities. 1964.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Daniel Alabi, Nicole Immorlica, and Adam Kalai. Unleashing linear optimizers for group-fair learning and optimization. In *Conference On Learning Theory*, pp. 2043–2066, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica*, 2016.

Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. *arXiv preprint arXiv:1902.03519*, 2019.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Ioana O Bercea, Martin Groß, Samir Khuller, Aounon Kumar, Clemens Rösner, Daniel R Schmidt, and Melanie Schmidt. On the cost of essentially fair clusterings. *arXiv preprint arXiv:1811.10319*, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Dimitri P Bertsekas. *Control of uncertain systems with a set-membership description of the uncertainty.* PhD thesis, Massachusetts Institute of Technology, 1971.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pp. 4349–4357, 2016.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 319–328. ACM, 2019.

Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *Advances in Neural Information Processing Systems*, pp. 5029–5037, 2017.

John M Danskin. *The theory of max-min and its application to weapons allocation problems*, volume 5. Springer Science & Business Media, 1967.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings. *Proceedings on privacy enhancing technologies*, 2015(1):92–112, 2015.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pp. 2791–2801, 2018.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pp. 119–133, 2018.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

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

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pp. 144–152. SIAM, 2016.

Hans Gebelein. Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941.

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005a.

Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *Journal of Machine Learning Research*, 6(Dec):2075–2129, 2005b.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pp. 520–524. Cambridge University Press, 1935.

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Faisal Kamiran and Toon Calders. Classification with no discrimination by preferential sampling. In *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*, pp. 1–6. Citeseer, 2010.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118, 2018.

Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

Maher Nouiehed, Maziar Sanjabi, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *arXiv preprint arXiv:1902.08297*, 2019.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 189–198. IEEE, 2018.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fair logistic regression: An adversarial perspective. *arXiv preprint arXiv:1903.03910*, 2019.

Clemens Rösner and Melanie Schmidt. Privacy preserving clustering with constraints. *arXiv preprint arXiv:1802.02497*, 2018.

Salvatore Ruggieri. Using t-closeness anonymity to control for non-discrimination. *Trans. Data Privacy*, 7(2):99–129, 2014.

Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854*, 2018.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2019.

Latanya Sweeney. Discrimination in online ad delivery. *arXiv preprint arXiv:1301.6822*, 2013.

Hans S Witsenhausen. On sequences of pairs of dependent random variables. *SIAM Journal on Applied Mathematics*, 28(1):100–113, 1975.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, 2018.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, 2018.

A APPENDIX: PROOF OF THEOREM 3.2

Proof. First of all, notice that since $\tilde{\mathbf{a}}$ is one-hot encoded version of a, any function $f : \{0, \ldots, c\} \mapsto \mathbb{R}$ can be equivalently represented as $f(\tilde{\mathbf{a}}) = \mathbf{u}^T \tilde{\mathbf{a}}$ for some $\mathbf{u} \in \mathbb{R}^c$. Therefore, according to the definition of Rényi correlation, we can write

$$\begin{aligned} \rho_R(a,b) = \max_{\mathbf{u},g} \quad & \mathbb{E}\left[(\mathbf{u}^T\tilde{\mathbf{a}})g(b)\right] \\ \text{s.t.} \quad & \mathbb{E}\left[(\mathbf{u}^T\tilde{\mathbf{a}})^2\right] \leq 1, \quad \mathbb{E}[\mathbf{u}^T\tilde{\mathbf{a}}] = 0 \\ & \mathbb{E}[g^2(b)] \leq 1, \quad \mathbb{E}[g(b)] = 0 \end{aligned}$$

Notice that since b is binary, there is a unique function $g(b) = \frac{b-q}{\sqrt{q(1-q)}}$ satisfying the constraints where $q \triangleq \mathbb{P}(b=1)$. Therefore, the above optimization problem can be written as

$$\begin{aligned} \rho_R(a,b) = \max_{\mathbf{u}} \quad & \mathbf{u}^T\mathbb{E}\left[\tilde{\mathbf{a}}g(b)\right] \\ \text{s.t.} \quad & \mathbf{u}^T\mathbb{E}\left[\tilde{\mathbf{a}}\tilde{\mathbf{a}}^T\right]\mathbf{u} \leq 1 \\ & \mathbf{u}^T\mathbb{E}[\tilde{\mathbf{a}}] = 0 \end{aligned}$$

The last constraint simply implies that $\mathbf{u}$ should be orthogonal to $\mathbf{p} \triangleq \mathbb{E}[\tilde{\mathbf{a}}]$. Equivalently, we can write $\mathbf{u} = \left(\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2}\right)\mathbf{v}$ for some $\mathbf{v} \in \mathbb{R}^c$. Thus, we can simplify the above optimization problem as

$$\begin{aligned} \rho_R(a,b) = \max_{\mathbf{v}} \quad & \mathbf{v}^T\left(\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2}\right)\mathbb{E}\left[\tilde{\mathbf{a}}g(b)\right] \\ \text{s.t.} \quad & \mathbf{v}^T\left(\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2}\right)\text{diag}(\mathbf{p})\left(\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2}\right)\mathbf{v} \leq 1, \end{aligned}$$

where in the constraint, we used the equality $\mathbb{E}\left[\tilde{\mathbf{a}}\tilde{\mathbf{a}}^T\right] = \text{diag}(\mathbf{p})$. Let us do the change of variable $\hat{\mathbf{v}} = \text{diag}(\sqrt{\mathbf{p}})(\mathbf{I} - \frac{\mathbf{p}\mathbf{p}^T}{\|\mathbf{p}\|^2})\mathbf{v}$. Then the above optimization can be simplified to

$$\begin{aligned} \rho_R(a,b) = \max_{\hat{\mathbf{v}}} \quad & \hat{\mathbf{v}}^T\text{diag}(1/\sqrt{\mathbf{p}})\mathbb{E}\left[\tilde{\mathbf{a}}g(b)\right] \\ \text{s.t.} \quad & \|\hat{\mathbf{v}}\| \leq 1. \end{aligned}$$

Clearly, this leads to

$$\begin{aligned} \rho_R^2(a,b) &= \left\|\text{diag}\left(\frac{1}{\sqrt{\mathbf{p}}}\right)\mathbb{E}\left[\tilde{\mathbf{a}}g(b)\right]\right\|^2 \\ &= \sum_{i=1}^{c}\frac{1}{\mathbb{P}(a=i)}\left(\mathbb{P}(a=i,b=1)\sqrt{\frac{1-q}{q}} - \mathbb{P}(a=i,b=0)\sqrt{\frac{q}{1-q}}\right)^2, \end{aligned} \tag{14}$$

where in the last equality we use the fact that $g(1) = \sqrt{\frac{1-q}{q}}$ and $g(0) = -\sqrt{\frac{q}{1-q}}$. Define $p_{i0} \triangleq \mathbb{P}(a=i, b=0)$ and $p_{i1} \triangleq \mathbb{P}(a=i, b=0)$, $p_i \triangleq \mathbb{P}(a=i) = p_{i0} + p_{i1}$. Then, using simple algebraic manipulations, we have that

$$\begin{aligned} \rho_R^2(a,b) &= \sum_{i=1}^{c}\frac{1}{p_i}\left(p_{i1}\sqrt{\frac{1-q}{q}} - p_{i0}\sqrt{\frac{q}{1-q}}\right)^2 \\ &= \sum_{i=1}^{c}\frac{(2p_{i1}(1-q) - 2p_{i0}q)^2}{4p_iq(1-q)} - \sum_{i=1}^{c}\frac{(p_{i0}-p_{i1})^2}{4p_iq(1-q)} + \sum_{i=1}^{c}\frac{(p_{i0}-p_{i1})^2}{4p_iq(1-q)} \\ &= \sum_{i=1}^{c}\frac{((3-2q)p_{i1} - (1+2q)p_{i0})((1-2q)p_{i1} + (1-2q)p_{i0})}{4p_iq(1-q)} + \sum_{i=1}^{c}\frac{(p_{i0}-p_{i1})^2}{4p_iq(1-q)} \\ &= \frac{1-2q}{4q(1-q)}((3-2q)q - (1+2q)(1-q)) + \sum_{i=1}^{c}\frac{(p_{i0}-p_{i1})^2}{4p_iq(1-q)} \\ &= 1 - \frac{1-\sum_{i=1}^{c}(p_{i0}-p_{i1})^2/p_i}{4q(1-q)} = 1 - \frac{\gamma}{q(1-q)}, \end{aligned}$$

where in the last equality we used the definition of γ and the optimal value of equation 5. $\square$

B PROOF OF THEOREM 4.2

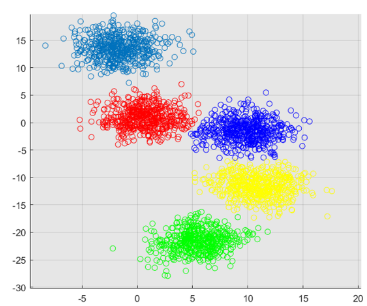
Proof. Define $g_B(\boldsymbol{\theta}) = \max_{\mathbf{w}} f_B(\boldsymbol{\theta}, \mathbf{w})$. Since the optimization problem $\max_{\mathbf{w}} f_B(\boldsymbol{\theta}, \mathbf{w})$ is strongly concave in $\mathbf{w}$, using Danskin's theorem (see Danskin (1967) and Bertsekas (1971)), we conclude that the function $g_B(\cdot)$ is differentiable. Moreover,

$$\nabla_{\boldsymbol{\theta}} g_B(\bar{\boldsymbol{\theta}}) = \nabla_{\boldsymbol{\theta}} f_B(\bar{\boldsymbol{\theta}}, \bar{\mathbf{w}})$$

where $\bar{\mathbf{w}} = \arg\max_{\mathbf{w}} f_B(\bar{\boldsymbol{\theta}}, \mathbf{w})$. Thus Algorithm 2 is in fact equivalent to the gradient descent algorithm applied to $g_B(\boldsymbol{\theta})$. Thus according to (Nesterov, 2018, Chapter 1), the algorithm finds a point with $\|\nabla g_B(\boldsymbol{\theta})\| \leq \epsilon$ in $\mathcal{O}(\epsilon^{-2})$ iterations. □

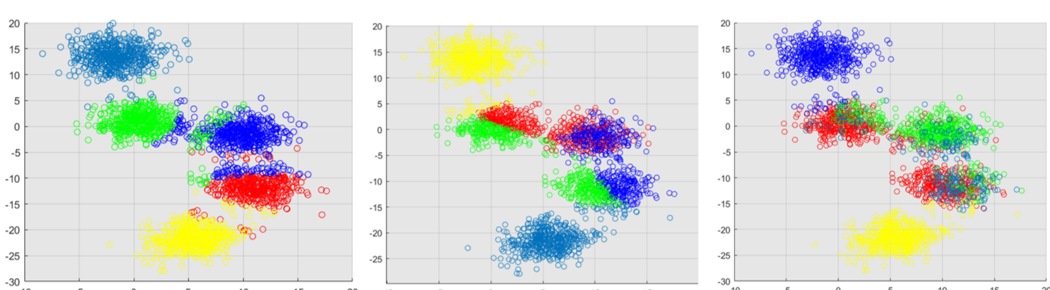
C RÉNYI FAIR K-MEANS

To illustrate the behavior of algorithm 3, we deployed a simple two-dimensional toy example. In this example we generated data by randomly selecting 5 center points and then randomly generating 500 points around each center according to a normal distribution with small enough variance. The data is shown in Figure 4 with different colors corresponding to different clusters. Moreover, we assign for each data point x_i a binary value $s_i \in \{0, 1\}$ that corresponds to the sensitive attribute. This assignment is also done randomly except for points generated around center 1 which are assigned a value of 1 and points generated around center 2 which are assigned a value of 0. Without imposing fairness, traditional K-means algorithm would group points generated around center 1 in one cluster regardless of the fact that they all belong to the same protected group. Similarly, points generated around center 2 will belong to the same cluster. However, as the experiment will show, our algorithm will distribute these points among various clusters to achieve balanced clusters. The clusters corresponding to centers 1 and 2 are the yellow and red clusters in Figure 4. It is evident from Figure 5 that increasing lambda, data points corresponding to centers 1 and 2 are now distributed among different clusters.



$\lambda = 0$

Figure 4: Applying K-means algorithm without fairness on the synthetic dataset.



$\lambda = 10$ $\lambda = 100$ $\lambda = 1000$

Figure 5: Applying fair K-means algorithm with different values of λ on the synthetic dataset.

C.1 UPDATING $\mathbf{w}$ AFTER UPDATING THE ASSIGNMENT OF EACH DATA POINT IN ALGORITHM 3

To understand the reasoning behind updating the vector of proportions $\mathbf{w}$ after updating each $\mathbf{a}_i$ which is the assignment of data point i, we discuss a simple one-dimensional counterexample. Consider the following four data points $X_1 = -5$, $X_2 = -4$, $X_3 = 4$, and $X_4 = 5$ with their corresponding sensitive attributes $S_1 = S_2 = 1$ and $S_3 = S_4 = 0$. Moreover, assume the following initial $\mathbf{A}^0$ and $\mathbf{C}^0$

$$\mathbf{A}^0 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{C}^0 = [-4.5, 4.5].$$

Hence, X_1 and X_2 which both have a sensitive attribute of 1 are assigned to cluster 1 with center $\mathbf{C}_1^0 = -4.5$ and X_3 and X_4 which both have a sensitive attribute of 0 are assigned to cluster 2 with center $\mathbf{C}_2^0 = 4.5$. Then $\mathbf{w}$ which is the current proportion of the privileged group in the clusters will be $\mathbf{w}^0 = [1, 0]$. Now, for sufficiently large λ if we update $\mathbf{A}$ according to Step 6 of Algorithm 3, we get the following new assignment

$$\mathbf{A}^1 = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{C}^1 = [4.5, -4.5], \quad \mathbf{w}^1 = [0, 1].$$

Hence, the points just switch their clusters. Then, performing another iteration will get us back to the initial setup and the algorithm will get stuck between these two states that are both not fair. To overcome this issue we update the proportions $\mathbf{w}$ after updating the assignment of each data point.

D DATASETS DESCRIPTION

In this section we introduce the datasets used in numerical experiment discussed in Section 6. All of these datasets are publicly available at UCI repository.

- **German Credit Dataset** [1]: German Credit dataset consists of 20 features (13 categorical and 7 numerical) regarding to social, and economic status of 1000 customers. The assigned task is to classify customers as good or bad credit risks. We chose first 800 customers as the training data, and last 200 customers as the test data.
- **Bank Dataset** [2]: Bank dataset contains the information of individuals contacted by a Portuguese bank institution. For the clustering task, we sampled 3 continuous features: Age, balance, and duration. The sensitive attribute is the marital status of the individuals. For the classification task we consider all 17 attributes (except martial status as the sensitive attribute). To evaluate the performance of the classifier, we split data into the training (32000 data points), and test set (13211 data points).
- **Adult Dataset** [3]: The adult dataset contains the census information of individuals including education, gender, capital-gain, and etc. For the clustering task, we chose 5 continuous features, and 10000 samples to cluster. The sensitive attribute of each individual is gender. For the classification task, we consider all 14 attributes (except gender and race as the sensitive attributes).

E SUPPLEMENTARY FIGURES

[1] `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`
[2] `https://archive.ics.uci.edu/ml/datasets/Bank%20Marketing`.
[3] `https://archive.ics.uci.edu/ml/datasets/adult`.

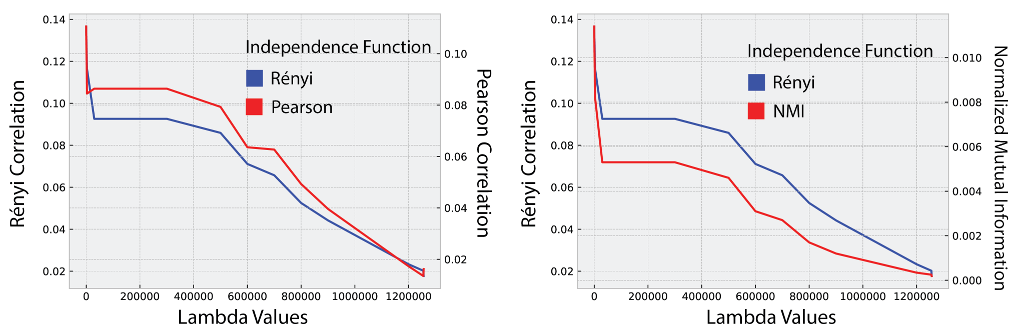


Figure 6: The relationship between Rényi correlation, Pearson correlation, and Normalized Mutual Information.