Fairness-Aware Learning for Continuous Attributes and Treatments

Jérémie Mary [1] **Clément Calauzènes** [1] **Noureddine El Karoui** [1 2]

Abstract

We address the problem of algorithmic fairness: ensuring that the outcome of a classifier is not biased towards certain values of sensitive variables such as age, race or gender. As common fairness metrics can be expressed as measures of (conditional) independence between variables, we propose to use the Rényi maximum correlation coefficient to generalize fairness measurement to continuous variables. We exploit Witsenhausen's characterization of the Rényi correlation coefficient to propose a differentiable implementation linked to f-divergences. This allows us to generalize fairness-aware learning to continuous variables by using a penalty that upper bounds this coefficient. Theses allows fairness to be extented to variables such as mixed ethnic groups or financial status without thresholds effects. This penalty can be estimated on mini-batches allowing to use deep nets. Experiments show favorable comparisons to state of the art on binary variables and prove the ability to protect continuous ones [1].

1. Introduction

Ensuring that sensitive information (e.g. knowledge about the ethnic group of an individual) does not "unfairly" influence the outcome of a learning algorithm is receiving increasing attention with the spread of AI tools in society. To achieve this goal, as discussed with more details in the related work section 3.1, there are three families of approaches: first modify a pretrained classifier while minimizing the impact on performance (Hardt et al., 2016; Pleiss et al., 2017), second enforce fairness during the training possibly at the cost of convexity (Zafar et al., 2017) and third modify data representation and use classical algorithms (Zemel et al., 2013; Donini et al., 2018). As formulated by (Hardt et al., 2016), the core ingredient of algorithmic fairness is the ability to estimate and guarantee (conditional) independence between two well chosen random variables – typically involving the decision made by the algorithm and the variables to protect and the "positive" outcome. We will call U and V these two random variables in the rest of this introduction. While in all generality U and V can be continuous variables – as an example a predicted probability or a variable such as time – most of the work in fairness has so far focused on protecting categorical variables. In this work we relax this assumption.

From an applied perspective this is desirable since it avoids to consider continuous values as predetermined "categorical bins" which are going to present thresholds effects in the learnt model. Theses thresholds make no real sense when considering age, ethnic proportions or a gender fluidity measure. Moreover, a smooth and continuous way to describe fairness constraints - a way that considers also the order of the elements (e.g. 10 yo $<$ 11 yo) - is important. As an example from the real world, (Daniels et al., 2000) pointed the financial status to be a sensitive variable for health care.

From a statistical point of view, given that the dependence between U and V can be arbitrarily complex, the measure of dependence is challenging. On one side of the spectrum – the empirical one – simple and tractable correlation coefficients have been introduced, such as Pearson's rho, Spearman's rank or Kendall's tau. Sadly, while such correlation coefficients are able to disprove independence, they are not able to prove it. They only express necessary conditions for independence. On the opposite side – the theoretical one – Gebelein (Gebelein, 1941) introduced the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient (HGR), which has the desirable property to prove independence when evaluated to zero (Rényi, 1959) but is computationally intractable in the general case. However, HGR is a measure of dependence which takes value in [0,1] and is independent of the marginals of U and V, which allows a regulator to use it with an absolute threshold as a fairness criterion.

Many authors have proposed principled ways to estimate this coefficient. Early approaches includes Witsenhausen's characterization (Witsenhausen, 1975) which provides a

[1] Criteo AI Lab, Paris, France [2] University of California, Berkeley, USA. Correspondence to: J. Mary <j.mary@criteo.com>, C. Calauzènes <c.calauzenes@criteo.com>, N. El Karoui <nkaroui@berkeley.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

[1] Code available at https://github.com/criteo-research/continuous-fairness

characterization by the second singular value of a well chosen matrix as discussed in Subsection 2.3. Later, (Breiman & Friedman, 1985a) used non-parametric estimates and effectively used the power method on the operator $U = P_U P_V$ – with P_V (resp. P_U) denotes the operator conditional expectation wrt V (resp. U). More recently (Lopez-Paz et al., 2013) exploited the largest canonical correlation between random non-linear projections of the respective empirical copula-transformations of U and V and showed excellent empirical behavior even for multi-dimensional variables, extending the work of (Bach & Jordan, 2003; Gretton et al., 2005; Reshef et al., 2011).

We first aim to take advantage of advances on measure of independence for algorithmic fairness. Simultaneously the rise of deep learning and the mood for *differentiable* programming advocates the usage of differentiable approximations with a nice first order behavior and a limited computational cost in order to be usable to penalize neural nets. In this work we derive a differentiable non parametric estimation based on Witsenhausen's characterization mixed with a Kernel Density Estimation (KDE). We demonstrate the empirical performance of this estimation and we tightly upper bound it by a quantity which is a f-divergence. The proposed bound is attained as soon as one of the two random variables (U or V) is binary valued. Note that f-divergences between a joint distribution and the product of its marginals are invariant under the action of invertible maps (Nelsen, 2010)) and can be used as measures of dependence.

As a second contribution we demonstrate that our upper bound on the HGR coefficient can used to penalize the learning of a model. It even empirically proves to perform well when estimated on mini-batches, allowing its use in conjunction with neural networks trained by stochastic gradient. Another key difference is that we are able to deal with sensitive variables which are continuous. Our approach also extends the work of Kamishima et al. (2011), who proposed to use an estimate of the Mutual Information (MI) as a penalty, but their estimation method is restricted to categorical variables. Moreover, even when extending the MI to the continuous case, we show that our regularizer yields both better models and less sensitivity to the value of hyper-parameters.

The rest of the paper is organized as follows: first we make explicit the link between measure of independence and different fairness criteria such as *Disparate Impact* and *Equalized Odds*. Secondly, we introduce our approximation of HGR which is able to deal with continuous variables and then use it to regularize losses for algorithmic fairness. We follow with an experimental section where we empirically demonstrate that our approximation is competitive with state of the art for dependency estimation and algorithmic fairness when the sensible attribute is categorical and that the minimization can be done using mini-batches of data if the dataset is large enough. Finally, we show that our approache generalizes to continuous sensitive attributes.

2. HGR as a Fairness Criterion

2.1. Overview of Fairness Criteria

Many notions of fairness are currently being investigated an there is not yet a consensus on which ones are the most appropriate (Hardt et al., 2016; Zafar et al., 2017; Dwork et al., 2012). Indeed, it is a choice that requires not only statistical and causal arguments, but also ethical ones. A common thread between these metrics is to rely on statistical independence. Fairness meets machine learning when one tries to build a prediction $\hat{Y}$ of some variable Y (ex: default of payment) based on some available information X (ex: credit card history); prediction that may be biased or unfair with respect to some sensitive attribute Z (ex: gender).

An initial notion proposed to measure fairness was the *demographic parity* of the prediction – i.e. whether $\mathbb{P}(\hat{Y} = 1|Z=1) = \mathbb{P}(\hat{Y}=1|Z=0)$ – which is often measured by the *disparate impact* criterion (Feldman et al., 2015):

$$\mathrm{DI} = \frac{\mathbb{P}(\hat{Y}=1|Z=0)}{\mathbb{P}(\hat{Y}=1|Z=1)}$$

This criterion is even part of the US Equal Employment Opportunity Commission recommendation (EEOC., 1979) which advocates that it should not be below 0.8 – also known as the 80% rule. While initially defined for binary variables, the *demographic parity* can be easily generalized to the requirement $\hat{Y} \perp\!\!\!\perp Z$, even when Z is non binary.

Demographic parity was criticized for ignoring confounding variables that may explain an already existing correlation in the data between Y and Z. For instance, a model selecting randomly 10% of men and choosing the best 10% women would be perfectly fair w.r.t. DI. To partially overcome these limitations, *Equalized Odds* was introduced by (Zafar et al., 2017; Hardt et al., 2016) as a measurement of whether $\mathbb{P}(\hat{Y} = 1|Z = 1, Y = y) = \mathbb{P}(\hat{Y} = 1|Z = 0, Y = y)$ for any y. The particular case of $y = 1$ only is referred to as *Equal Opportunity* (EO) and commonly measured by the difference of EO:

$$\mathrm{DEO} = \mathbb{P}(\hat{Y}=1|Z=1, Y=1) - \mathbb{P}(\hat{Y}=1|Z=0, Y=1)$$

Again, similarly to *Demographic Parity*, *Equal Opportunity* can be represented equivalently by a notion of independence, namely $\hat{Y} \perp\!\!\!\perp Z|Y$. Some other notions of fairness have been developed, such as the notion of *Calibration* that tries to ensure that the prediction $\hat{Y}$ is not more accurate for a protected group than for another – namely that $Y \perp\!\!\!\perp Z|\hat{Y}$ – but we refer the interested reader to (Hardt et al., 2016; Zafar et al., 2017; Kilbertus et al., 2017) for more insights on the notions of fairness that have been proposed over time.

While much effort has been invested in finding new definitions of fairness to cover different possible social biases, the associated statistical measures have remained restricted to binary values of Y and Z. The prediction $\hat{Y}$ is often also considered binary, but most definitions extend to continuous values for $\hat{Y}$ by using divergences between conditional distributions as suggested in Dwork et al. (2012). In the following, we first use the interpretation of the different notions of fairness as (potentially conditional) independence measures to generalize their measurement to multi-variate and continuous cases.

2.2. Measuring Independence: the HGR coefficient

Measuring dependence and testing for independence has a long history in statistical learning and we turn to this literature to propose new ways of measuring fairness and proposing methods for fair machine learning. The central question is therefore how to measure that two random variables, not necessarily of the same dimension, discrete or continuous are independent. A natural object is the so-called maximum correlation coefficient (a.k.a. Rényi correlation). See (Gebelein, 1941; Rényi, 1959)

Definition 2.1 (Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient)**.** *Given two random variables $U \in \mathcal{U}$ and $V \in \mathcal{V}$, the Hirschfeld-Gebelein-Rényi Maximum Correlation Coefficient is defined as follow,*

$$\text{HGR}(U,V) = \sup_{f,g} \rho(f(U), g(V)) \tag{1}$$

where ρ is the Pearson's correlation coefficient and f, g are (measurable) functions with $\mathbf{E}\left[f^2(U)\right], \mathbf{E}\left[g^2(V)\right] < \infty$.

As a measure of (in)dependence, HGR has several desirable properties: it is in $[0,1]$, is 0 if and only if V and U are independent, is 1 if U and V are deterministically linked by an equation. It is obviously invariant under invertible transforms of U and/or V. Furthermore, it admits a simple geometric interpretation : it is the cosine of the angle between the space of square integrable functions of U and those of V (viewed as Hilbert subspaces in the Hilbert space of square integrable functions of (U,V)) ((Bickel et al., 1998) Thm. 2 p. 438, (Breiman & Friedman, 1985b)). It plays a central role in the analysis of transformations/early representation learning through the ACE algorithm (Breiman & Friedman, 1985b) that tries to find the best non-trivial representation/transformation of V and U so as to minimize their square distance. We refer to (Breiman & Friedman, 1985b) and (Pregibon & Vardi, 1985) for an early but still relevant discussion of HGR as a practical tool in data analysis.

In our context, one of its main interest is that it can handle continuous and discrete variables and allows to generalize in a principled way many notions of fairness.

Computable Approximation of HGR As we just explained, HGR is an abstract measure that needs to be approximated to be computed on a dataset. If we limit ourselves to f and g linear, HGR is of course just the canonical correlation between the two variables. Naturally, this classical remark makes it natural to approximate HGR by requiring f and g to belong to Reproducing Kernel Hilbert spaces and this has been done efficiently (Lopez-Paz et al., 2013) under the name Randomized Dependency Coefficient (RDC). We will describe below our own and different approach as well as links with classical measures in information theory.

Finally, we note that there has been recent interest in new measures of independence such as Brownian distance covariance ((Székely & Rizzo, 2009)), which can be used to test independence. We refer to (Bell, 1962) for a comparison of HGR and measures derived from the mutual information.

2.3. On Witsenhausen's Characterization

HGR has also been of interest in the information theory literature. (Witsenhausen, 1975) shows the following.

Theorem 2.2 (Witsenhausen)**.** *Suppose U and V are discrete random variables and consider the matrix*

$$Q(u,v) = \frac{\pi(u,v)}{\sqrt{\pi_U(u)}\sqrt{\pi_V(v)}} \; , \tag{2}$$

where $\pi(u,v)$ is the joint distribution of u and v, and π_U and π_V are the corresponding marginals. Then

$$\text{HGR}(U,V) = \sigma_2(Q) \; , \tag{3}$$

where σ_2 is the second largest singular value of a matrix.

As noted in his paper, the result extends to the case of continuous random variables, provided some compactness assumptions on Q, viewed then as the kernel of a linear operator on $L^2(d\pi_U d\pi_V)$. To avoid cumbersome notations, we sometime write $\pi(u)$ instead of $\pi_U(u)$.

Corollary 2.2.1. *Under the assumptions of Theorem 2.2,*

$$\text{HGR}^2 \leq \sum_{u,v} \frac{\pi(u,v)^2}{\pi_U(u)\pi_V(v)} - 1 = \chi^2(\pi_{U,V}, \pi_U \otimes \pi_V) \; .$$

where χ^2 is the f-divergence with $f(t) = t^2 - 1$.

The results extend for continuous variables with sum(s) replaced by integral(s), provided $\int \frac{\pi(u,v)^2}{\pi_U(u)\pi_V(v)} dudv < \infty$.

If one of the two variables U or V is binary, the inequality is an equality.

Proof. In the matrix case, if we denote by $\sigma_i(Q)$ the decreasingly ordered singular values of Q, i.e. $\sigma_1(Q)$ is the largest singular value, we know that

$$\sigma_1^2(Q) + \sigma_2^2(Q) \leq \sum_i \sigma_i^2(Q) = \sum_{u,v} Q(u,v)^2 \; ,$$

as the sum of squared singular values is the square of the Hilbert-Schmidt norm of the matrix. (Witsenhausen, 1975) establish that if π_V and π_U denote vector of marginal distributions, $Q\sqrt{\pi_V} = \sqrt{\pi_U}$ and $Q'\sqrt{\pi_U} = \sqrt{\pi_V}$. This follows from the fact that $\sum_u \pi(u,v) = \pi_V(v)$ and $\sum_v \pi(u,v) = \pi_U(u)$. Thus if $\sqrt{\pi_V}$ is an eigenvector of $Q'Q$ associated to the eigenvalue 1. Through operator theoretic arguments, (Witsenhausen, 1975) shows that the largest singular value of Q is always 1. The inequality above yields

$$\text{HGR}^2(Q) = \sigma_2^2(Q) \leq \sum_{u,v} Q(u,v)^2 - 1\ .$$

The results extend to the case of compact operators through standard functional analytic arguments.

Let $f(t) = t^2 - 1$. The associated f-divergence is $D_f(\pi(u,v)||\pi(u)\pi(v)) = \sum_{u,v}([\pi(u,v)/(\pi(u)\pi(v)]^2 - 1)\pi(u)\pi(v)$. This is exactly our upper bound, as $\sum_{u,v}\pi(u)\pi(v) = 1$. This divergence is called the χ^2 divergence. It is known to be an upper bound of the KL divergence when both are defined because on $\mathbb{R}^+$ we have $t\ln(t) - t + 1 \leq (t-1)^2$.

If U (resp V) is binary then Q has two rows (resp. columns), so when one variable is binary the matrix Q is of rank at most two. As $\sigma_1(Q) = 1$ the other (possibly) non-zero singular value of Q is $\sigma_2(Q) = \text{HGR}(Q)$, hence $1 + \sigma_2^2(Q) = \sum_{u,v} Q(u,v)^2$. The inequality is an equality here. The continuous case is treated in the same way. □

Much of our approach below consists in using the Corollary above to penalize our machine learning algorithms by HGR, using the exact χ^2 divergence representation in the common case where the sensitive variable is binary. However, the approach brings flexibility to the problem when the protected attributes are not binary. In the case of a variable like ethnic statistics of a district, which is continuous, we could either use a threshold to make it binary (and use existing techniques) or simply use its continuous form. It is clear that the thresholding approach is somewhat undesirable.

Note on invariance : because HGR correlation is invariant under 1-1 transformation, we can always assume that the continuous variable has Unif[0,1] marginals. This essentially amounts to working with a copula version of the data. Calling $\widetilde{V}$ this transformed version of V, we see that if U is binary, we can rewrite HGR as

$$\text{HGR}^2 = \sum_u \frac{1}{\pi_U(u)}\mathbf{E}\left[\pi^2(u,\widetilde{V})\right] - 1\ .$$

This formulation gives us a simple way to evaluate HGR^2 by subsampling on $\widetilde{V}$. Figure 1 shows that this formulation leads to an efficient estimation of the level of dependence between real variables when joint law is estimated using a gaussian kernel density estimation described in Sec. 4.1.

2.4. Statistical estimation of HGR

Witsenhausen's characterization gives the possibility to use different estimations methods of HGR depending on the fairness problem we are facing, whether both $\hat{Y}$ and Z are binary (or finite), one is continuous and the other is binary or both are continuous. We present here the two main practical ways to estimate HGR empirically: for discrete variables and for continuous variables.

Categorical variables : In this setting the $\hat{Q}$ matrix is straightforward and finite. Yet the estimation of the singular values of a stochastic matrix from a finite sample is known to be difficult. More precisely, the bootstrap is known to be problematic for eigenvalues of symmetric matrices that have multiplicity higher than one (see e.g. (Eaton & Tyler, 1991)). In essence the difficulty comes from the fact that we cannot do a Taylor expansion in this setting and the statistic is not smooth. One possible approach is of course to use an m-out-of-n bootstrap (see e.g. (Bickel et al., 1997) or (Eaton & Tyler, 1991) in a close but slightly different context). But in practice the choice of m is known to be difficult while in our setting, the following Lemma grants us consistency.

Lemma 2.3. *Suppose that the two variables U and V are categorical and hence take only finitely many values. Then the elements of the empirical Witsenhausen matrix $\widehat{Q}(u,v) = \frac{\widehat{\pi}(u,v)}{\sqrt{\widehat{\pi}_U(u)\widehat{\pi}_V(v)}}$ are jointly asymptotically normal as $n \to \infty$.*

We have the conservative bound

$$|\text{HGR}(\widehat{Q}) - \text{HGR}(Q)|^2 \leq \frac{1}{n} trace\left(E'E\right)\ ,$$

where $E = \sqrt{n}(\widehat{Q} - Q)$. The bootstrap can be used to estimate the statistical properties of this upper bound and give conservative confidence intervals for $\text{HGR}(Q)$.

The proof can be found in the Appendix, **Note :** We show in the proof of the Lemma that, if the second singular value of Q has multiplicity 1, then $\text{HGR}(\hat{Q})$ is asymptotically normal and $\sqrt{n}$ consistent for $\text{HGR}(Q)$. In that case, the bootstrap can be used to accurately estimate its statistical variability. The limit theorem above is needed to assess the statistical variability of our estimator and e.g. perform hypothesis tests.

Continuous Variables In the continuous case, keeping in mind that we want to be able to derive a penalization scheme for learning, we propose to use a Gaussian KDE to build an estimate $\hat{\pi}(u,v)$ of $\pi(u,v)$. To avoid introducing new hyper-parameters, we are using the classical Silverman's rule (Silverman, 1986) for setting the bandwidth of the kernel on the normalized data. Numerically in order to marginalize efficiently we compute the estimation of the density on a regular square grid.

A note on conditional independence. We just saw how to derive estimators of HGR to measure the dependence between two variables, which allows to extend the measurement of *demographic parity* to the continuous case. As we saw in section 2.1, other notions of fairness like *equalized odds* and *calibration* rely on conditional independence. We can note that $Q(u|w, v|w) = \frac{\pi(u,v,w)}{\sqrt{\pi(u,w)\pi(v,w)}}$. It means that we can simply include the conditioning variable in the estimation and apply the same method as before to obtain $Q(u|w, v|w)$ which allows to compute HGR as a function of the conditioning. Then, we can use a functional norm $||.||$ to use $||\text{HGR}(u|., v|.)||$ as a conditional independence criterion to derive fairness metrics.

3. Fairness Aware Learning

In the previous section, we advocated for the use of HGR to derive fairness evaluation criteria. It is then natural to ask whether this method can be used to derive penalization schemes in order to enforce fairness during the learning phase. In order to avoid cumbersome notation and discussion, we focus on the *equalized odds* setting, but similar learning schemes can be derived for the other fairness settings. In particular, we provide a corresponding set of experiments in the Appendix for the setting of *demographic parity*. As a reminder, given some class of hypothesis $\mathcal{H}$, we wish to learn $\hat{Y} = h(X)$ with $h \in \mathcal{H}$, that regresses Y over some variables X while providing an *equalized odds* guarantee w.r.t. a sensitive attribute Z – i.e. ensuring that we control $\text{HGR}|_\infty \triangleq ||\text{HGR}(\hat{Y}|Y = y, Z|Y = y)||_\infty$. Given an expected loss $\mathcal{L}$, a class of function $\mathcal{H}$ and a fairness tolerance $\varepsilon > 0$, we want to solve the following problem:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{L}(h, X, Y) \qquad \text{subject to HGR}|_\infty \leq \varepsilon \tag{4}$$

Unfortunately, in the general case, estimating HGR requires to compute the singular values of $Q(u, v)$. Hence, we use the upper bound on $\text{HGR}(u, v)$ provided in Corollary 2.2.1 by $\chi^2(u, v)$. Fortunately, all methods proposed in Section 2.4 to estimate HGR are based on estimating Q and hence also valid for the χ^2-divergence as $\chi^2(u, v) = \int Q^2(u, v)dudv - 1$. Then we can relax the constraint $\text{HGR}|_\infty \leq \varepsilon$ in (4) in several directions to obtain a tractable penalty. First, we relax HGR to the χ^2-divergence. Then, we use a density estimate $\hat{\pi}(\hat{y}, z, y)$ of the "true" density $\pi(\hat{y}, z, y)$ using the best-suited estimation technique from Section 2.4. Finally, since z is continuous and we have finite sample size to optimize, we also relax the infinite norm to a L_1 norm. Then, defining,

$$\chi^2|_1 = \left\|\chi^2\left(\hat{\pi}(\hat{y}|y, z|y), \hat{\pi}(\hat{y}|y) \otimes \hat{\pi}(z|y)\right)\right\|_1 \tag{5}$$

We obtain a penalized learning objective:

$$\underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \mathcal{L}(h, X, Y) + \lambda \chi^2|_1 \tag{6}$$

Our method can easily be generalized to any other f-divergence beside the χ^2, leading to different penalties. In section 4.3.1 additional experiments in appendix, we also implement a penalty based on the KL-divergence, which corresponds to penalizing with the MI.

3.1. State of the Art

Learning fair models is a topic of growing interest as machine learning is becoming a very common tool for insurance companies, health system, law... So far, research focused on the binary case ($\hat{Y}$, Y and Z binary).In this setting, it is possible to calibrate *a posteriori* the model predicting $\hat{Y}$ based on Z and Y in order to satisfy a DEO constraint for instance, at the expense of losing some accuracy, either by re-weighting the probabilities output by the model (Calders & Verwer, 2010) or by adapting the classification threshold (Hardt et al., 2016). In order to embed fairness in the learning procedure to potentially find better trade-offs between accuracy and fairness than what can be achieved *a posteriori*, Menon & Williamson (2018) integrated the re-weighting at learning time and proposed a cost-sensitive classification approach for fairness. However, such approaches are deeply linked to the binary nature of the variables of interest.

Another approach is to add a fairness constraint (or equivalently a penalty) to the optimization objective during the learning. Two types of constraints have been proposed so far. A first type of methods follows from the seminal paper of Dwork et al. (2012). A simplified instantiation of their method for *equalized odds* is to constrain the conditional distributions $\pi(\hat{Y}|Z = z, Y = y)$ to be close for any y provided some distance $\mathcal{D}$ between distributions (or equivalently close to their marginals). Bechavod & Ligett (2017) advocates for replacing $\mathcal{D}$ by the distance between the first moments of the conditional distributions. This idea was generalized (Donini et al., 2018) to the distance between the expectation of a function of $\hat{Y}$ which allows the use of higher moments of the conditional distribution. The second type of approach corresponds to ideas that are linked to our approach. Kamishima et al. (2011) is a simplification of our approach, where the f-divergence used is the KL but the estimation of the penalty is specific to the binary case. Lastly, (Zafar et al., 2017) propose a constraint aiming at minimizing the conditional covariance between $\hat{Y}$ and Z, which corresponds to removing linear correlations only, while HGR can handle more complex cases. This last point is especially important when aiming at using such penalties with deep neural networks that definitely have the capacity to over-fit a penalty only accounting for linear dependencies.

Finally, there have been a line of work aiming to propose measures of fairness for regression (i.e., the setting in which the target variable is continuous) (Komiyama et al., 2018; Speicher et al., 2018). Theses are classically using a covariance (which is a subcase of Renyi) or another less general metric such as Gini. Also this relies on properties of linear systems - possibly in a kernel space - while we regularize a deep net and are not restricted to demographic parity.

4. Experiments

In order to support empirically the different claims we made above, we provide several experiments. We first show that our approximation of HGR is competitive with RDC (Lopez-Paz et al., 2013) when testing independence on real valued variables. Second, in the context of training a fair classifier – i.e. we want the binary outcome of a classifier to be fair w.r.t. a binary feature Z – we check that our approximation can be used to regularize a classifier in order to enforce fairness with results comparable to the state of the art. Here we show that when the dataset has a few thousand data points, we can ensure fairness of its estimated probability of positive classification. Finally, we show that we can protect the output of a classifier w.r.t. a continuous sensitive attribute.

4.1. Witsenhausen's characterization with KDE

We first reproduce some of the experiments proposed by (Lopez-Paz et al., 2013) for RDC. The power of a dependence measure is defined as its ability to discern between dependent and independent samples that share equal marginal forms and is expressed as a probability. Here we compare our HGR-KDE estimation against RDC as a measure of nonlinear dependence. We replicated 7 bivariate association patterns, given in Figure 1. For each of the association patterns F, 500 repetitions of $n = 500$ samples were generated, in which we sample from $X \sim$ unif$[0, 1]$ getting $(x_i, F(x_i))$ tuples. Next, we regenerated the input variable independently from $Y \sim$ unif$[0, 1]$, to generate independent versions of each sample with equal marginals $(Y, F(X))$. Figure 1 shows the power for the discussed non-linear dependence measures as the standard deviation of some zero-mean Gaussian additive noise increases from 0 to 3. We observe the power of χ^2 to be very similar to the HGR-KDE estimation while the performance is better than RDC on circle, linear and sinus associations while for quadratic and cubic association RDC performs slightly better. Empirically on one dimensional data, our χ^2 estimation is competitive with RDC while its simple and differentiable form allows us to compute it at a very reasonable cost. On a recent laptop computing HGR-KDE with 500 tuples takes 2.0ms using our pytorch implementation while the published numpy code for RDC requires 4.6ms (average done on 1000 runs).

Remark: we used the Silverman's rule (Silverman, 1986) to set the KDE bandwidth to $\delta = (n\frac{d+2}{4})^{\frac{-1}{d+4}}$, where n is the number of samples on which the KDE is estimated and d is the dimension of the joint distribution to estimate. All estimations are done on a normalized version of the data.

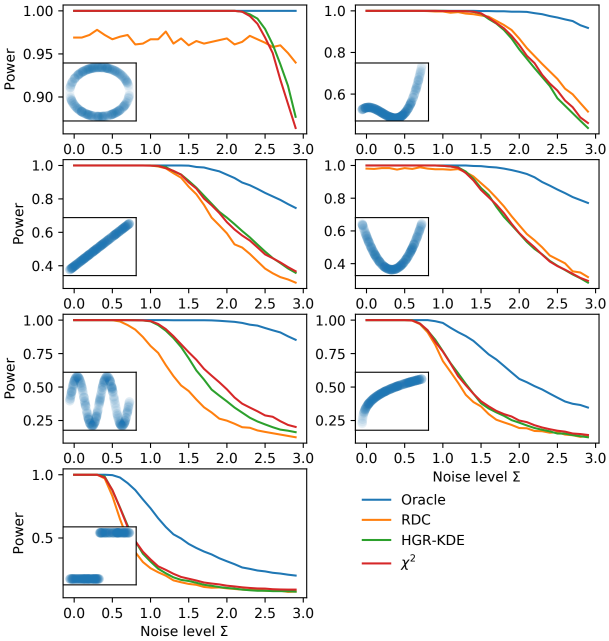

Figure 1. Power of dependency identification – higher is better – w.r.t the noise level Σ for different association given in the subplots. Given two independent random uniform variables X, Y on $[0; 1]$ and a association F we try to separate data with distribution $(X, F(X) + \varepsilon)$ w.r.t data with distribution $(Y, F(X) + \varepsilon)$ using 500 samples and $\varepsilon \sim \mathcal{N}(0, \Sigma)$. Oracle is using knowledge of F.

4.2. Fairness with Y and Z binary valued.

In this experiment we address the use of different independence measures to penalize the training of a non linear neural network in order to train a classifier such that a binary sensitive information Z (e.g. knowledge about the ethnic group of an individual or Sex in this experiment) does not *unfairly* influence an outcome $\hat{Y}$. In order to prove that this regularization is competitive with the state of the art on binary variables we reproduce the experiments from (Donini et al., 2018). They propose to use 5 publicly available datasets: Arrhythmia, COMPAS, Adult, German, and Drug. A description of the datasets as well as the variable to protect is provided in the supplementary material of (Donini et al., 2018). These datasets are from UCI and the proposed task yields a DEO higher than 0.1 when the used classifier is a SVM. Reported results for their method and baselines is from their work except FERM which has avalaible implentation and were we report the best result between results in the paper and a rerun with our preprocessing. As a preprocessing step, we one hot encode all categorical variables and normalize the numeric entries. As the goal is to maintain a good accuracy while having a smaller DEO we use (6). In the binary case the penalty term collapses to an estimation of $\chi^2(\hat{Y}, Z|Y = 1)$ where χ^2 is equal to HGR^2. Here for $\hat{Y}$

	Arrhythmia		COMPAS[2]		Adult		German		Drug	
Method	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Naïve SVM	0.75 ± 0.04	0.11 ± 0.03	0.72 ± 0.01	0.14 ± 0.02	0.80	0.09	0.74 ± 0.05	0.12 ± 0.05	0.81 ± 0.02	0.22 ± 0.04
SVM	0.71 ± 0.05	0.10 ± 0.03	0.73 ± 0.01	0.11 ± 0.02	0.79	0.08	0.74 ± 0.03	0.10 ± 0.06	0.81 ± 0.02	0.22 ± 0.03
FERM	0.75 ± 0.05	0.05 ± 0.02	0.96 ± 0.01	0.09 ± 0.02	0.77	0.01	0.73 ± 0.04	0.05 ± 0.03	0.79 ± 0.03	0.10 ± 0.05
NN	0.74 ± 0.07	0.19 ± 0.14	0.97 ± 0.00	0.01 ± 0.00	0.84	0.14	0.74 ± 0.04	0.47 ± 0.19	0.79 ± 0.03	0.15 ± 0.16
NN + χ^2	0.75 ± 0.06	0.15 ± 0.09	0.96 ± 0.00	0.00 ± 0.00	0.83	0.03	0.73 ± 0.03	0.25 ± 0.14	0.78 ± 0.05	0.00 ± 0.00

Table 1. ± standard on 10 runs when protected variable is outside of dataset, more results in appendix. Most of Values in this table are from (Donini et al., 2018). NN is not using any fairness aware technique while $NN + \chi^2$ is using the regularizer described in section 4.2.

we use the probability of $Y = 1$ estimated by the network to build the estimate. This is referred as NN + χ^2 in Tab. 1.

Structure of the Neural Net and learning. We used a simple neural net NN for these experiments: two hidden layers (first layer is from 30 to 100 neurons depending on the size of the data set, the second being 20 neurons smaller than the first one). Non linearities are SELU (Klambauer et al., 2017). Loss is crossentropy, gradient is Adam (Kingma & Ba, 2014) and learning rate is from values from $10^{-2}, 10^{-4}, 3 \cdot 10^{-4}$. Batch size is chosen from $\{8, 16, 32, 64, 128\}$. To avoid estimation issues for the KDE – which occurs especially when $Y = 1$ is rare – we always estimate the χ^2 penalty over a separate batch of size in 128. λ is set to $4 *$ Rényi batch size/batch size. Also remark that we use χ^2 which is the square of HGR because the values of the gradient close to 0 are numerically more stable.

Remarks: Performance is bad on the Arrythmia and German datasets where we are not able to significantly reduce the DEO. Arrythmia is made of 451 examples while German has 1,000. Theses sizes are probably too small for our proposal when regularized loss is minimized thought gradient descent on a neural network. About Arrythmia and German, the absence of publication of the exact pre-processing is problematic. E.g. running Domini's code on those datasets with minimal pre-processing (one hot + normalization) yields a DEO on test of 0.25 which is higher than the scores we obtain while accuracies are of resp. of 0.77 and 0.72 (similar to ours). Their paper reports much better results we are not able to reproduce and we choose to report their values in the table 1. When the dataset is larger not knowing the pre-processing method is less important since the net we use can learn the representation. On datasets which contains a few thousand of samples, our proposed regularizer lead to a very competitive ”fair” classifier.

4.3. Fully Continuous Case: Criminality Rates

The last experiment highlights one of the main contributions of this paper: extending fairness-aware learning to protecting a *continuous* prediction $\hat{Y}$ w.r.t. a *continuous* protected variable Z given a continuous Y. The task here is to predict the number of violent crimes per population for US cities, while protecting race information (ratio of an ethnic group in the population) to avoid biasing police checks depending on the ethnic characteristics of the neighborhood [3].

Following Section 3, we consider an *equalized odds* objective and we evaluate the different algorithms in terms of $\text{HGR}|_\infty$[4] All the experiments are a result of a 10-fold cross-validation for which each fold is averaged over 20 random initializations (as the objective is non-convex).

As state-of-the-art algorithms are not handling the continuous case, the only way to compare our results to theirs is to apply them by binarizing the continuous variables. We choose to use Bechavod & Ligett (2017) as a baseline – denoted $L_2^{\hat{Y}|Z,Y}$ – as it can handle continuous $\hat{Y}$ and only require to binarize Z and Y, which makes it more competitive than other methods that would require to discretize all of them to be applied even optimizing the thresholds.

For our method, we propose to study empirically two penalties derived from our framework: a $\chi^2|_1$ penalty as described in (6) and a variant with the KL replacing the χ^2, denoted $\text{KL}|_1$. The KL penalty estimated with KDE is a generalization of the prejudice removal penalty (Kamishima et al., 2011) proposed in the binary case. As explained in Section 2.4, for $\chi^2|_1$ and $\text{KL}|_1$, we use a Gaussian KDE with bandwidth set following Silverman's rule with $d = 3$.

4.3.1. PENALIZED BATCH LINEAR REGRESSION

First, we consider a linear regression (LR) penalized with the three different fairness penalties. This allows us to observe how these penalty terms are behaving when estimated with full batches of data. All the methods are optimized using L-BFGS. Comparisons are reported on Fig. 2.

We can observe on Figure 2 that the $\chi^2|_1$ penalty manages to provide the best trade-off between fairness and MSE, even though the best runs for the other two penalties are close. That is another strength of the $\chi^2|_1$: in full batch, it is more robust to the regularization parameter setup than the other ones. Note that some values of regularization are not visible on the graph for $\text{KL}|_1$ and $L_2^{\hat{Y}|Z,Y}$, the MSE being very degraded. The case of $L_2^{\hat{Y}|Z,Y}$ is even worse as not only is it sensitive to the regularization parameter but it also required to tune the discretization.

[3] The dataset is available here: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`

[4] The same results evaluated in $\text{KL}|_\infty$ are available in Appendix and show very similar results.

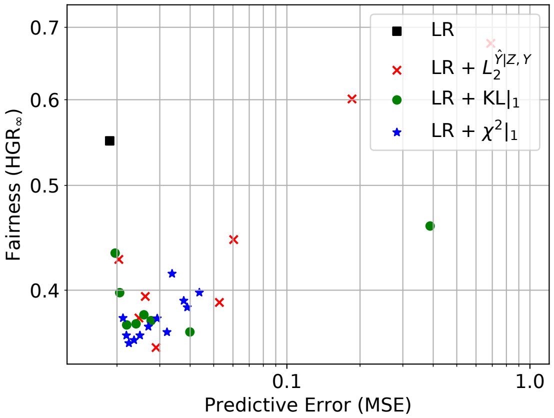

Figure 2. Equalized odds with Linear Regression: Compromise between predictive performance (MSE) and fairness (HGR$|_\infty$) that are reached by the different algorithms. The Pareto front for each algorithm is obtained by varying the hyper-parameter controlling the level of fairness (regularization parameter) from 2^{-4} to 2^6. For KL$|_1$ and $L_2^{\hat{Y}|Z,Y}$ some points are out of the graph on the right.

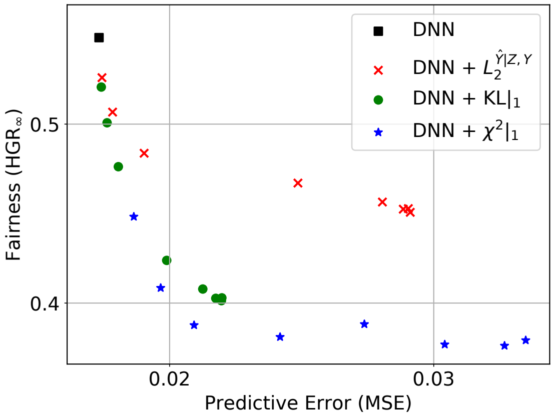

Figure 3. Equalized odds with DNN: Compromise between predictive performance (MSE) and fairness (HGR$|_\infty$) that are reached by the different algorithms. The Pareto front for each algorithm is obtained by varying the hyper-parameter controlling the level of fairness (regularization parameter) from 2^{-4} to 2^6.

Sample complexity is a bottleneck for improving fairness. We can observe on all methods the fairness measure first decreases with the strength of regularization, then increases again. This signals over-fitting of the fairness penalty (see Appendix to compare with the same graph on the training set). It is not surprising in the *equalized odds* setting, as we aim at making independent two continuous variables conditionally to a third one, and this for any value of this third one. Given the size of our training set, it is to be expected that very low values of HGR$|_\infty$ can't be reached in test. Corroborating this claim, in *Demographic Parity* – a statistically easier task as there is no conditioning – the values of HGR$(\hat{Y}, Z)$ reached in test are much lower and the over-fitting of the fairness penalty is smaller.

4.3.2. Penalized Mini-batch Learning with DNN

Finally, we emphasize the ability of our method to work with mini-batch optimization, a crucial need to make it compatible with the learning of deep neural networks models. Now $h \in \mathcal{H}$ is a DNN and the objective (6) is optimized with a stochastic optimizer (Adam) with mini-batches of size $n = 200$. The bandwidth of the KDE is still set using the same heuristic, except n is now the size of a mini-batch.

On Fig. 3 we observe that, independently from the penalty, DNNs are able to improve fairness at a lower price than linear models in terms of MSE thanks to their larger capacity: even with high regularization values, the MSE only increases to 0.03. This supports the requirement for a fairness penalty to be compatible with the learning of deep models. Then, we can observe that the baseline ($L_2^{\hat{Y}|Z,Y}$ penalty) suffers from the mini-batching due to the binarization of Z and Y, making it unsuited for deep learning. On the contrary, the $\chi^2|_1$ and KL$|_1$ penalties prove to work smoothly with mini-batched stochastic optimization used for deep learning and achieve satisfying compromises between fairness and MSE. Remark in appendix we have variants of Fig 2 and 3 with HGR being replaced with MI.

5. Conclusion

Thanks to HGR, we have unified and extended previous frameworks to continuous sensitive information for algorithmic fairness from both evaluation and learning point of views. First, we proposed a principled way to derive evaluation measures for fairness objectives that can be written as conditional independence. Then, we provided the corresponding derivation for the learning step. Finally we empirically show the performance of our approach on a series of problems (continuous or not) and the adaptability to deep learning models. An interesting question left for future work is whether the non-parametric density estimation done with KDE could be replaced by a parametric estimation to improve the scaling of the method and reduce the variance in the context of mini-batching.

6. Acknowledgements

N. El Karoui gratefully acknowledges support from grant NSF DMS 1510172.

References

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, March 2003. ISSN 1532-4435. doi: 10.1162/153244303768966085.

Bechavod, Y. and Ligett, K. Learning fair classifiers: A regularization-inspired approach. *arXiv pre-print*, abs/1707.00044, 2017.

Bell, C. B. Mutual information and maximal correlation as measures of dependence. *Ann. Math. Statist.*, 33(2): 587–595, 06 1962. doi: 10.1214/aoms/1177704583.

Bickel, P. J., Götze, F., and van Zwet, W. R. Resampling fewer than n observations: Gains, losses and remedies for losses. *Statistica Sinica*, 7(1):1–31, 1997.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. *Efficient and adaptive estimation for semiparametric models*. Springer-Verlag, New York, 1998. ISBN 0-387-98473-9. Reprint of the 1993 original.

Breiman, L. and Friedman, J. Estimating optimal transformations for multiple regression and correlation: Rejoinder. *Journal of The American Statistical Association*, 80:580–598, 09 1985a. doi: 10.1080/01621459.1985.10478157.

Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.*, 80(391):580–619, 1985b. ISSN 0162-1459. With discussion and with a reply by the authors.

Calders, T. and Verwer, S. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, Sep 2010. ISSN 1573-756X. doi: 10.1007/s10618-010-0190-x.

Daniels, N., Bryant, J., Castano, R., Gomez-Dantes, O., s khan, K., and Pannarunothai, S. Benchmarks of fairness for health care reform: A policy tool for developing countries. *Bulletin of the World Health Organization*, 78:740–50, 02 2000. doi: 10.1590/S0042-96862000000600006.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M. Empirical risk minimization under fairness constraints. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 2796–2806. Curran Associates, Inc., 2018.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pp. 214–226, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1115-1. doi: 10.1145/2090236.2090255.

Eaton, M. L. and Tyler, D. E. On wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.*, 19(1):260–271, 03 1991. doi: 10.1214/aos/1176347980.

EEOC., T. U. Uniform guidelines on employee selection procedures. March 1979.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. 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.

Gebelein, H. Das statistische problem der korrelation als variations- und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik*, 21(6):364–379, 1941. doi: 10.1002/zamm.19410210604.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *Proceedings of the 16th International Conference on Algorithmic Learning Theory*, ALT'05, pp. 63–77, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-29242-X, 978-3-540-29242-5. doi: 10.1007/11564089_7.

Hardt, M., Price, E., , and Srebro, N. Equality of opportunity in supervised learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3315–3323. Curran Associates, Inc., 2016.

Horn, R. A. and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986. ISBN 0-521-30586-1.

Kamishima, T., Akaho, S., and Sakuma, J. Fairness-aware learning through regularization approach. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*, ICDMW '11, pp. 643–650, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-0-7695-4409-0. doi: 10.1109/ICDMW.2011.83.

Kato, T. *Perturbation theory for linear operators; 2nd ed.* Grundlehren Math. Wiss. Springer, Berlin, 1976.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30*, pp. 656–666. Curran Associates, Inc., December 2017.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 971–980. Curran Associates, Inc., 2017.

Komiyama, J., Takeda, A., Honda, J., and Shimao, H. Nonconvex optimization for regression with fairness constraints. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2737–2746, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/komiyama18a.html.

Lopez-Paz, D., Hennig, P., and Schölkopf, B. The randomized dependence coefficient. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pp. 1–9, USA, 2013. Curran Associates Inc.

Menon, A. K. and Williamson, R. C. The cost of fairness in binary classification. In Friedler, S. A. and Wilson, C. (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 107–118, New York, NY, USA, 23–24 Feb 2018. PMLR.

Nelsen, R. B. *An Introduction to Copulas*. Springer Publishing Company, Incorporated, 2010. ISBN 1441921095, 9781441921093.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5680–5689. Curran Associates, Inc., 2017.

Pregibon, D. and Vardi, Y. Comment. *Journal of the American Statistical Association*, 80(391):598–601, 1985. doi: 10.1080/01621459.1985.10478158.

Rényi, A. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungaricae*, 10(3-4):441–451, sep 1959. doi: 10.1007/bf02024507.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, December 2011. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1205438.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., and Zafar, M. B. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pp. 2239–2248, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5552-0.

Székely, G. J. and Rizzo, M. L. Brownian distance covariance. *Annals of Applied Statistics*, 3(4):1236–1265, oct 2009. ISSN 19326157. doi: 10.1214/09-AOAS312.

Vaart, A. W. v. d. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.

Witsenhausen, H. S. On sequences of pairs of dependent random variables. *SIAM J. Appl. Math.*, 28:100–113, 1975. ISSN 0036-1399. doi: 10.1137/0128010.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *arXiv*, März 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In Dasgupta, S. and McAllester, D. (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

A. Asymptotic normality of HGR

We have the following lemma.

Lemma A.1. *Suppose that the two variables X and Y are categorical and hence can take finitely many values. Then the elements of the empirical Witsenhausen matrix $\widehat{Q}(u,v) = \frac{\widehat{\pi}(u,v)}{\sqrt{\widehat{\pi}(u)\widehat{\pi}(v)}}$ are jointly asymptotically normal as $n \to \infty$.*

Furthermore, if the second singular value of Q has multiplicity 1, then $\mathrm{HGR}(\widehat{Q})$ is asymptotically normal and $\sqrt{n}$ consistent for $\mathrm{HGR}(Q)$. In that case, the bootstrap can be used to accurately estimate its statistical variability.

Finally, we have

$$\left(\mathrm{HGR}(\widehat{Q}) - \mathrm{HGR}(Q)\right)^2 \leq \frac{1}{n} trace\,(E'E) \ ,$$

with $E = \sqrt{n}(\widehat{Q} - Q)$ and the statistical properties of this upper bound can be obtained by bootstrapping. This yields a conservative asymptotic confidence interval for $\mathrm{HGR}(\widehat{Q})$.

Note : as the statistic in our upper bound is a simply modification of the χ^2 statistic for independence, we recall that the usual rule of thumb for applicability of the χ^2 test in the null case is that we have 5 observations per cell. In this context, it seems that the simplest choice of bootstrap to use would be a parametric bootstrap using multinomial sampling with the observed $\widehat{\pi}(x,y)$ if $\mathrm{HGR}(\widehat{Q})$ appears to be separated from the other singular values of $\widehat{Q}$. We note that our upper bound on HGR^2 is the standard χ^2-statistic for testing independence minus 1. So in case the singular values of $\widehat{Q}$ are not sufficiently separated, we can use the upper bound mentioned in the Lemma.

Proof. We note that $\hat{\pi}(u,v)$ are the realization of a Multinomial random vector with parameters $(\{\pi(u,v)\}_{u,v}, n)$. It is well known (see (Vaart, 1998), Chapter 17) that

$$\sqrt{n}\left(\frac{\hat{\pi}(u,v) - \pi(u,v)}{\sqrt{\pi(u,v)}}\right)_{u,v} \Rightarrow \mathcal{N}(0, \mathrm{id} - \sqrt{\pi}\sqrt{\pi}') \ .$$

In other words, the entries of the matrix $\widehat{\pi}(u,v)$ are asymptotically jointly normal with

$$n\mathrm{cov}(\hat{\pi}(u,v), \hat{\pi}(u',v')) = -\hat{\pi}(u,v), \hat{\pi}(u',v') \\ +\pi(u,v)^2\mathbf{1}[u=u', v=v'] \ .$$

Let us call $R(u,v) = \pi(u,v)/(\pi(u)\pi(v))$,$W_n(u,v) = \frac{\hat{\pi}(u,v)-\pi(u,v)}{\sqrt{\pi(u,v)}}$, and $[\widehat{\pi}_U(u)/\pi_U(u)][\widehat{\pi}_V(v)/\pi_V(v)] = Z_n(u,v)$. We can clearly write

$$\widehat{Q}(u,v) = \frac{\widehat{\pi}(u,v) - \pi(u,v) + \pi(u,v)}{\sqrt{\pi_U(u)\pi_V(v)}} Z_n^{-1/2}$$
$$= \frac{\widehat{\pi}(u,v) - \pi(u,v)}{\sqrt{\pi(u,v)}} \sqrt{R(u,v)} Z^{-1/2} + Q(u,v) Z_n^{-1/2}$$
$$= W_n(u,v)\sqrt{R(u,v)} Z_n^{-1/2} + Q(u,v) Z_n^{-1/2} \ .$$

Now the previous result on multinomial distributions guarantees that $\sqrt{n}W_n$ is asymptotically $\mathcal{N}(0,1)$. The same result implies that $\widehat{\pi}_U(u)$ is also asymptotically normal as a finite sum of asymptotically Gaussian random variable and the same result applies to $\widehat{\pi}_V(v)$. This normality is obviously joint since we are just looking at various linear transformations of the vector $\left(\frac{\hat{\pi}(u,v)-\pi(u,v)}{\sqrt{\pi(u,v)}}\right)_{u,v}$. The δ-method ((Vaart, 1998), p. 26) guarantees that $\sqrt{n}(W_n, (Z_n^{-1/2} - 1)$ is jointly asymptotically normal. The same reasoning applies to $\sqrt{n}(W_n, (Z_n - 1)$. We conclude using Slutsky's lemma ((Vaart, 1998), p. 11). Finally we note that the arguments above give that

$$\sqrt{n}\left(\widehat{Q}(u,v) - Q(u,v)\right) = \sqrt{n}W_n(u,v)\sqrt{R(u,v)} \\ -\frac{Q(u,v)}{2}\sqrt{n}(Z_n - 1) + O_P(n^{-1/2}) \ ,$$

in the standard probabilistic Landau notation of Section 2.2 in (Vaart, 1998). In other words the entries of the empirical Witsenhausen matrix are $\sqrt{n}$ consistent for the population version. And the matrix of errors

$$E = \sqrt{n}\left(\widehat{Q} - Q\right)$$

has all its entries jointly asymptotically Gaussian. The jointly Gaussian nature of its entries follows the same reasoning as above and remarking that its vectorized form, vec(E) is essentially a linear tranformation of the asymptotically Gaussian vector $\sqrt{n}\left(\frac{\hat{\pi}(u,v)-\pi(u,v)}{\sqrt{\pi(u,v)}}\right)_{u,v}$

Now recall that $\mathrm{HGR}^2 = \lambda_2(Q'Q)$, where $'$ denotes transposition. It is well known (see (Kato, 1976)) that if $\lambda_k(M)$, the k-th eigenvalue of a symmetric matrix M is simple, and if $\widehat{M} = M + t\Delta$, where Δ has bounded operator norm and t goes to zero, we have

$$\lambda_k(\widehat{M}) = \lambda_k(M) + t\phi_k'\Delta\phi_k + O(t^2) \ ,$$

where ϕ_k is an eigenvector associated with $\lambda_k(M)$ (of course ϕ_k is unique up to sign). In our case, we see that

$$\widehat{Q}'\widehat{Q} = Q'Q + \frac{1}{\sqrt{n}}(Q'E + E'Q) + \frac{1}{n}E'E \ .$$

Because of our results on E above, we see that when $\lambda_2(Q'Q)$ is simple, we have

$$\sqrt{n}\left[\lambda_2(\widehat{Q}'\widehat{Q}) - \lambda_2(Q'Q)\right] = 2\phi_2' Q E \phi_2 + O_P(n^{-1/2}) \ .$$

Writing the singular value decomposition of Q as

$$Q = \sum_i \sigma_i(Q)\psi_i\phi_i' \ ,$$

we have $\phi_2' Q' = \sigma_2(Q)\psi_2'$, so that

$$\sqrt{n}(\text{HGR}^2(\widehat{Q}) - \text{HGR}^2(Q)) = 2\sqrt{n}\text{HGR}(Q)\psi_2' E\phi_2 + O_P(n^{-1/2}) \ .$$

Applying the δ-method with $f(x) = \sqrt{x}$, we finally get

$$\sqrt{n}(\text{HGR}(\widehat{Q}) - \text{HGR}(Q)) = \sqrt{n}\psi_2' E\phi_2 + O_P(n^{-1/2}) \ .$$

As this last quantity is a linear form in an asymptotically Gaussian vector, it is asymptotically Gaussian. This shows the result.

When the eigenvalues are not separated, or to be conservative, we can use the standard bound ((Horn & Johnson, 1986), p. 419)

$$\left(\sigma_2(\widehat{Q}) - \sigma_2(Q)\right)^2 \leq \frac{1}{n}\text{trace}\,(E'E) \ ,$$

and the statistical properties of E to get conservative confidence intervals. We note that in practice the statistical properties of E could be obtained by using a parametric bootstrap to estimate E and the statistical properties of trace $(E'E)$. □

B. Additional Experimental Results

We provide here some additional results to complete the experiments of Section 4.3.

B.1. Penalty Over-Fitting

First, we provide a complement of Figures 2 and 3 but evaluated on both the training set and the test set to highlight the over-fitting phenomenon.

B.2. Evaluation in Mutual Information

Then, we provide a complement of Figures 2 and 3 evaluated in $\text{MI}|_\infty$ instead of $\text{HGR}|_\infty$. The observations to make are quite similar to the one made in the paper. Surprisingly, $\chi^2|_1$ stays better than $\text{KL}|_1$ at optimizing for the $\text{MI}|_\infty$.

B.3. Demographic Parity

Finally, we reproduce the same experiment but in *demographic parity*, meaning that we are trying to enforce $\hat{Y} \perp\!\!\!\perp Z$. Everything simplifies as an independence is easier to measure than a conditional one. The penalty becomes simply the χ^2 instead of $\chi^2|_1$ and the evaluation becomes HGR instead of $\text{HGR}|_\infty$. As this task is statistically simpler than a conditional independence, all the methods manage to improve the fairness much more as they are not over-fitting as much.

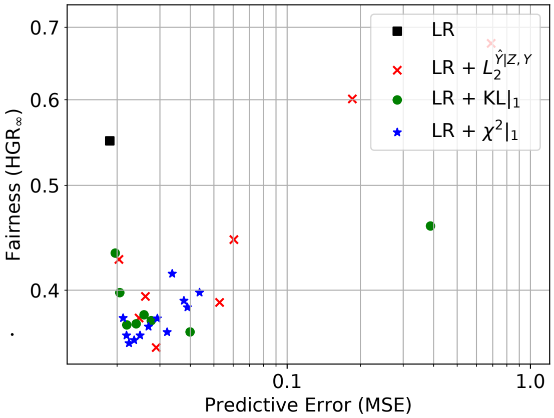

Figure 4. Equal Opportunity with Linear Regression on Test Set: Compromise between predictive performance (MSE) and fairness ($\text{HGR}|_\infty$).

	Arrhythmia		COMPAS[5]		Adult		German		Drug	
Method	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Variable to protect Z not inside the features used for prediction										
Naïve Lin. SVM	$0.75{\pm}0.04$	$0.11{\pm}0.03$	$0.73{\pm}0.01$	$0.13{\pm}0.02$	0.78	0.10	$0.71{\pm}0.06$	$0.16{\pm}0.04$	$0.79{\pm}0.02$	$0.25{\pm}0.03$
Lin. SVM	$0.71{\pm}0.05$	$0.10{\pm}0.03$	$0.72{\pm}0.01$	$0.12{\pm}0.02$	0.78	0.09	$0.69{\pm}0.04$	$0.11{\pm}0.10$	$0.79{\pm}0.02$	$0.25{\pm}0.04$
Zafar	$0.67{\pm}0.03$	$0.05{\pm}0.02$	$0.69{\pm}0.01$	$0.10{\pm}0.08$	0.76	0.05	$0.62{\pm}0.09$	$0.13{\pm}0.10$	$0.66{\pm}0.03$	$0.06{\pm}0.06$
Lin. FERM	$0.75{\pm}0.05$	$0.05{\pm}0.02$	$0.73{\pm}0.01$	$0.07{\pm}0.02$	0.75	0.01	$0.69{\pm}0.04$	$0.06{\pm}0.03$	$0.79{\pm}0.02$	$0.10{\pm}0.06$
Naïve SVM	$0.75{\pm}0.04$	$0.11{\pm}0.03$	$0.72{\pm}0.01$	$0.14{\pm}0.02$	0.80	0.09	$0.74{\pm}0.05$	$0.12{\pm}0.05$	$0.81{\pm}0.02$	$0.22{\pm}0.04$
SVM	$0.71{\pm}0.05$	$0.10{\pm}0.03$	$0.73{\pm}0.01$	$0.11{\pm}0.02$	0.79	0.08	$0.74{\pm}0.03$	$0.10{\pm}0.06$	$0.81{\pm}0.02$	$0.22{\pm}0.03$
FERM	$0.75{\pm}0.05$	$0.05{\pm}0.02$	$0.96{\pm}0.01$	$0.09{\pm}0.02$	0.77	0.01	$0.73{\pm}0.04$	$0.05{\pm}0.03$	$0.79{\pm}0.03$	$0.10{\pm}0.05$
NN	$0.74{\pm}0.07$	$0.19{\pm}0.14$	$0.97{\pm}0.00$	$0.01{\pm}0.00$	0.84	0.14	$0.74{\pm}0.04$	$0.47{\pm}0.19$	$0.79{\pm}0.03$	$0.15{\pm}0.16$
NN + χ^2	$0.75{\pm}0.06$	$0.15{\pm}0.09$	$0.96{\pm}0.00$	$0.00{\pm}0.00$	0.83	0.03	$0.73{\pm}0.03$	$0.25{\pm}0.14$	$0.78{\pm}0.05$	$0.00{\pm}0.00$
Variable to protect Z inside the features used for prediction										
Naïve Lin. SVM	$0.79{\pm}0.06$	$0.14{\pm}0.03$	$0.76{\pm}0.01$	$0.17{\pm}0.02$	0.81	0.14	$0.71{\pm}0.06$	$0.17{\pm}0.05$	$0.81{\pm}0.02$	$0.44{\pm}0.03$
Lin. SVM	$0.78{\pm}0.07$	$0.13{\pm}0.04$	$0.75{\pm}0.01$	$0.15{\pm}0.02$	0.80	0.13	$0.69{\pm}0.04$	$0.11{\pm}0.10$	$0.81{\pm}0.02$	$0.41{\pm}0.06$
Hardt Lin.	$0.74{\pm}0.06$	$0.07{\pm}0.04$	$0.67{\pm}0.03$	$0.21{\pm}0.09$	0.80	0.10	$0.61{\pm}0.15$	$0.15{\pm}0.13$	$0.77{\pm}0.02$	$0.22{\pm}0.09$
Zafar	$0.71{\pm}0.03$	$0.03{\pm}0.02$	$0.69{\pm}0.02$	$0.10{\pm}0.06$	0.78	0.05	$0.62{\pm}0.09$	$0.13{\pm}0.11$	$0.69{\pm}0.03$	$0.02{\pm}0.07$
Lin. FERM	$0.79{\pm}0.07$	$0.04{\pm}0.03$	$0.76{\pm}0.01$	$0.04{\pm}0.03$	0.77	0.01	$0.69{\pm}0.04$	$0.05{\pm}0.03$	$0.79{\pm}0.02$	$0.05{\pm}0.03$
Naïve SVM	$0.79{\pm}0.06$	$0.14{\pm}0.04$	$0.76{\pm}0.01$	$0.18{\pm}0.02$	0.84	0.18	$0.74{\pm}0.05$	$0.12{\pm}0.05$	$0.82{\pm}0.02$	$0.45{\pm}0.04$
SVM	$0.78{\pm}0.06$	$0.13{\pm}0.04$	$0.73{\pm}0.01$	$0.14{\pm}0.02$	0.82	0.14	$0.74{\pm}0.03$	$0.10{\pm}0.06$	$0.81{\pm}0.02$	$0.38{\pm}0.03$
Hardt	$0.74{\pm}0.06$	$0.07{\pm}0.04$	$0.71{\pm}0.01$	$0.08{\pm}0.01$	0.82	0.11	$0.71{\pm}0.03$	$0.11{\pm}0.18$	$0.75{\pm}0.11$	$0.14{\pm}0.08$
FERM	$0.79{\pm}0.09$	$0.03{\pm}0.02$	$0.96{\pm}0.01$	$0.09{\pm}0.02$	0.81	0.01	$0.73{\pm}0.04$	$0.05{\pm}0.03$	$0.80{\pm}0.03$	$0.07{\pm}0.05$
NN	$0.79{\pm}0.08$	$0.15{\pm}0.09$	$0.97{\pm}0.00$	$0.01{\pm}0.00$	0.84	0.13	$0.76{\pm}0.03$	$0.46{\pm}0.04$	$0.78{\pm}0.04$	$0.07{\pm}0.04$
NN + χ^2	$0.79{\pm}0.08$	$0.16{\pm}0.14$	$0.96{\pm}0.00$	$0.00{\pm}0.00$	0.83	0.08	$0.72{\pm}0.03$	$0.21{\pm}0.15$	$0.80{\pm}0.04$	$0.00{\pm}0.01$

Table 2. Results (average ± standard deviation averaged on 10 runs, when a fixed test set is not provided) for all the datasets, concerning accuracy (ACC) and DEO . Values in this table are from (Donini et al., 2018) except for the values highlighted with a blue background which are ours. NN is not using any fairness aware technique while $NN + \chi^2$ is using the regularizer described in section 4.2.

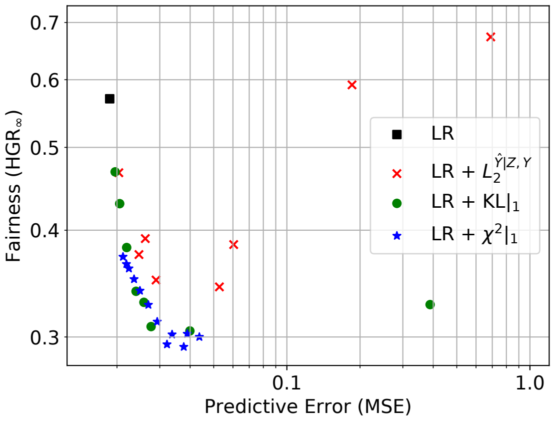

Figure 5. Equal Opportunity with Linear Regression on Training Set: Compromise between predictive performance (MSE) and fairness (HGR$|_\infty$).

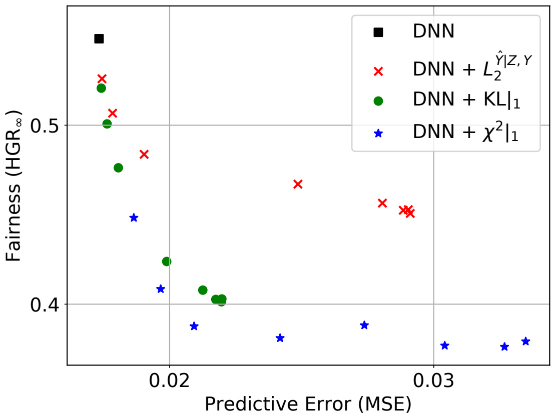

Figure 6. Equal Opportunity with DNN on Test Set: Compromise between predictive performance (MSE) and fairness (HGR$|_\infty$).

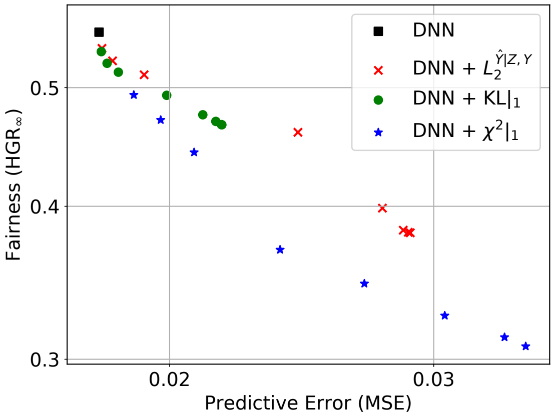

Figure 7. Equal Opportunity with DNN on Training Set: Compromise between predictive performance (MSE) and fairness (HGR$|_\infty$).

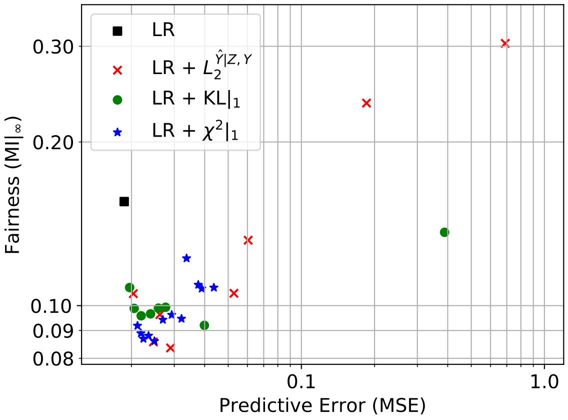

Figure 8. Equal Opportunity with DNN on Test Set: Compromise between predictive performance (MSE) and fairness (MI$|_\infty$).

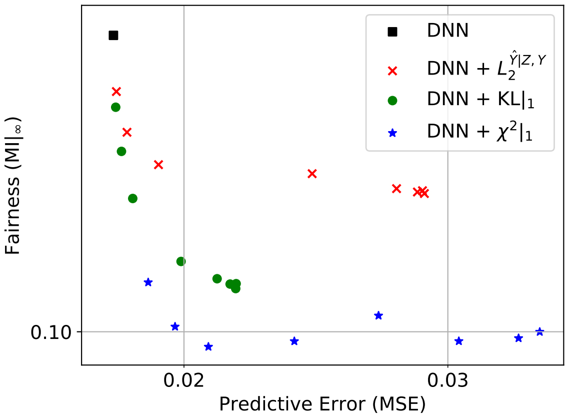

Figure 9. Equal Opportunity with DNN on Training Set: Compromise between predictive performance (MSE) and fairness (MI$|_\infty$).

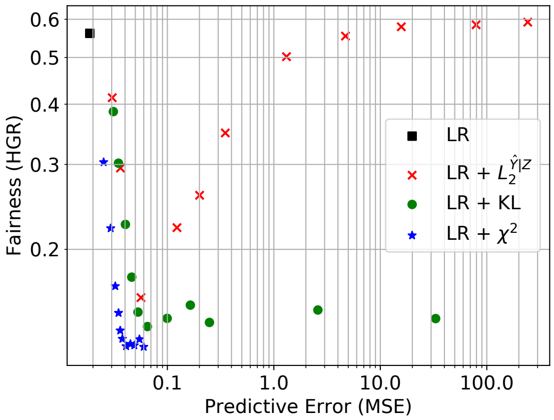

Figure 10. Demographic Parity with Linear Regression on Test Set: Compromise between predictive performance (MSE) and fairness (HGR).

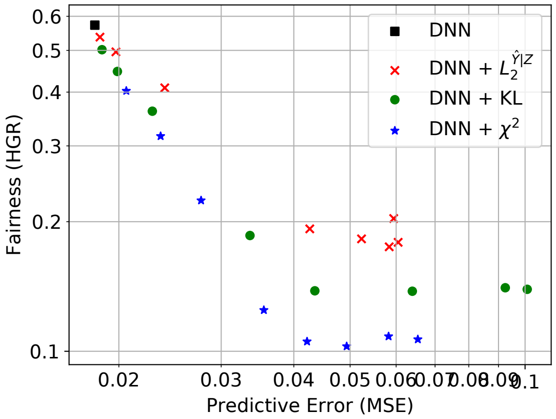

Figure 11. Demographic Parity with DNN on Test Set: Compromise between predictive performance (MSE) and fairness (HGR).