Conditionally Independent Data Generation

Kartik Ahuja[1] **Prasanna Sattigeri**[2] **Karthikeyan Shanmugam**[2] **Dennis Wei**[2]
Karthikeyan Natesan Ramamurthy[2] **Murat Kocaoglu**[3]

[1]Mila, Université de Montréal, Montreal, QC, Canada
[2]IBM Research, Yorktown Heights, NY, USA
[3]School of Electrical And Computer Engineering, Purdue University, West Lafayette, IN, USA

Abstract

Conditional independence (CI) is a fundamental concept with wide applications in machine learning and causal inference. Although the problems of testing CI and estimating divergences have been extensively studied, the complementary problem of generating data that satisfies CI has received much less attention. A special case of the generation problem is to produce conditionally independent predictions. Given samples from an input data distribution, we formulate the problem of generating samples from a distribution that is close to the input distribution and satisfies CI. We establish a characterization of CI in terms of a general divergence identity. Based on one version of this identity, an architecture is proposed that leverages the capabilities of generative adversarial networks (GANs) to enforce CI in an end-to-end differentiable manner. As one illustration of the problem formulation and architecture, we consider applications to notions of fairness that can be written as CIs, specifically equalized odds and conditional statistical parity. We demonstrate conditionally independent prediction that trades off adherence to fairness criteria against classification accuracy.

1 INTRODUCTION

Conditional independence (CI) is a fundamental probabilistic notion that has applications in causal inference and machine learning. In causal inference, CI tests are used to efficiently narrow down the space of causal graphs compatible with the given data, not only in observational but also in interventional settings where data from experiments are available [Yang et al., 2018]. In machine learning, CI tests are used as a non-parametric method for feature selection [Tsamardinos et al., 2003].

Due to its widespread uses, CI *testing* has been extensively studied in computer science, statistics, and information theory, as we discuss in Section 1.1. However, much less attention has been paid to the complementary problem of *generating* data with a desired CI, which commonly manifests as modifying a given dataset for which the CI is not satisfied. We pay particular attention to the case where the two variables that should be conditionally independent are an outcome variable and sensitive attributes. In this case, the generation problem becomes conditionally independent "fair" prediction.

The canonical problem of interest is as follows: Given samples of (X, Y, Z, W) drawn from $p(x, y, z, w)$, how can we generate samples from a distribution $\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w})$ such that: (a) $\tilde{X} \perp\!\!\!\perp \tilde{Y}|\tilde{Z}$ and (b) p and $\tilde{p}$ are close in some appropriate distance measure? In the case of fair prediction, Y is an "unfair" label and X are sensitive attributes. Here W are extra variables that do not participate in the CI expression, but could be important for reducing distance between p and $\tilde{p}$. This is because W could have information about (say) Y that is not captured by other variables. To address (a), we seek an approximate version of the conditional independence such that $\tilde{p}(\tilde{y}|\tilde{x}, \tilde{z})$ is close to $\tilde{p}(\tilde{y}|\tilde{z})$ in terms of a suitable distance/divergence measure. This is a key problem we address in this paper. The solution is non-trivial because the CI constraint is only on a subset of variables (X, Y, Z) while one needs to match p and $\tilde{p}$ in all the variables (X, Y, Z, W).

Although generation with CI constraints is a new problem, if one considers existing ideas in the testing literature, then enforcing CI could involve obtaining samples from a *perfect conditional sampler* for $p(y|z)$ (perhaps using a pre-trained conditional generator). We discuss some natural strategies that use this in Section 1.0.1. Some of these run into other roadblocks besides the difficulty of perfect conditional sampling when Z is high dimensional.

Our central idea rests on a general characterization of CI in terms of equality between two divergences that involve sam-

Accepted for the 37th Conference on Uncertainty in Artificial Intelligence (UAI 2021).

ples from p, $\tilde{p}$, and an *imperfect sampler* $q(y|z) \neq p(y|z)$. For bounded variables Y, the only requirement is that $q(y|z)$ has support overlap with $p(y|z)$ and this can be ensured by a uniform sampler on the bounded domain. We identify two key properties of the divergences, separability and strict convexity, that allow this result to be proven for a large class of divergences including Jensen-Shannon divergence, f-divergences, and Bregman divergences.

We develop a neural network architecture for approximate CI data generation. This is based on a special case of the characterization above for Jensen-Shannon divergence. We recall the standard GAN (Generative adversarial networks) architecture of Goodfellow [2016]. A discriminator is a parameterized function that computes an approximate distance measure between two distributions from their samples. When the discriminator's parameters are optimized, its output is a proxy for the distance measure. A generative model transforms a white noise input to produce samples from a distribution of interest. An adversarial game against the discriminator forces the generator to produce the desired distribution. In this context, our architecture (see Figure 1) has two additional discriminators (corresponding to the two divergences) and access to an *imperfect* sampler to enforce CI, apart from the standard components used to enforce closeness between p and $\tilde{p}$. Notably, it *eliminates* any need for additional pre-trained perfect generative models. The resulting CI-enforcing GAN enables a trade-off between how much the CI statement is enforced and how close the generated data is to the original dataset.

There are several potential applications of conditionally independent data generation. In this paper, we explore applications to fairness in machine learning, where many proposed criteria can be written as CI statements (we mention another application in Section 5). We focus on two criteria: 1) *equalized odds* (EO) [Hardt et al., 2016, Zafar et al., 2017], which requires CI between a predicted outcome $\hat{Y}$ and protected attribute S given the true outcome Y, and 2) *conditional statistical parity* (CSP) [Kamiran et al., 2013, Corbett-Davies et al., 2017], a generalization of statistical parity that requires CI of $\hat{Y}$ and S conditioned on a set of admissible variables A that are considered legitimate factors accounting for dependence between $\hat{Y}$ and S. We specialize the proposed CI-enforcing GAN architecture to these two criteria. Using the well-known Adult income dataset, our approach results in varying degrees of adherence to these criteria by tuning a hyperparameter, without unduly sacrificing classification accuracy. In the case of EO, for which there are many existing solutions, these results can be regarded as a proof of concept that the proposed CI generation method works. For CSP on the other hand, many fewer solutions exist and our contribution is more significant, being (as far as we are aware) the first to handle multiple admissible variables without having to enumerate all their values.

Our contributions: We proceed from general theory to specific applications, as summarized below:

1. We establish a general characterization of CI in terms of an identity that holds for a large class of divergences satisfying separability and strict convexity properties. This does not require access to samples from a conditional generator for $p(y|z)$.
2. Based on the Jensen-Shannon version of the identity, we propose an end-to-end differentiable GAN-based architecture for the problem of generating samples from a distribution that approximately satisfies a desired CI statement while remaining close to a given data distribution.
3. As an illustration of the utility of the architecture, we explore applications to fair classification in which predictions are generated to trade off between classification accuracy on the original dataset and the criteria of equalized odds or conditional statistical parity.

1.0.1 Key Technical Issue

We discuss some approaches that invariably rely on a pre-trained *perfect* conditional generative model (or a sampler along with a trained classifier if Y is categorical) to sample $\tilde{Y}$ from $p(y|z)$. A straightforward approach for enforcing CI is to try to replace the original Y samples by $\tilde{Y}$ sampled from $p(y|z)$ using the perfect sampler and substitute this for Y to obtain $\tilde{p}$. This will ensure CI in the subset $\tilde{X}, \tilde{Y}, \tilde{Z}$, after marginalizing over W. However, this approach generates $\tilde{Y}$ only from Z which is sub-optimal in terms of distance between p and $\tilde{p}$ since W could capture additional information about Y that is not captured by other variables and this information need not be sacrificed necessarily to impose CI. Another related solution would be to construct a reference distribution $p_r(x, y, z)$ such that p_r is the conditionally independent version of p over X, Y, Z, i.e. $p_r(x, y, z) = p(x, z)p(y|z)$, and then contrast $\tilde{p}$ with p_r using another discriminator to compute a distance between $\tilde{p}$ and p_r. The main drawback is that training a perfect conditional generator is difficult when Z is high dimensional and continuous. This issue has been recognized in the CI testing literature [Berrett et al., 2020]. In this work, we propose a different function that enforces CI only needing access to an *imperfect reference sampler* $q(y|z) \neq p(y|z)$. For bounded variables Y, only the support of $q(y|z)$ must overlap with $p(y|z)$ and this can even be ensured by a uniform sampler on the bounded domain.

1.1 RELATED WORK

To the best of our knowledge, the current work is unique in tackling the *generation* of data satisfying a CI statement in a differentiable manner. Below we discuss methods for *testing* CI and estimating divergences, some of which are not differentiable.

Conditional Mutual Information Estimation: Estimating conditional mutual information is a clear approach to testing CI [Póczos and Schneider, 2012] since two random variables are (conditionally) independent if and only if their (conditional) mutual information is zero. Estimation has traditionally been done by estimating multiple entropy terms using kernel density estimates [Gao et al., 2016]. Recently, Belghazi et al. [2018] proposed variational lower bounds for this task. In Gao et al. [2018], a very general principle for estimating divergence measures was introduced based on these variational lower bounds. Hash-based techniques for divergence estimation have also been used [Noshad et al., 2019]. However, these estimators are either not differentiable or they provide only a lower bound using a differentiable model. Our technique circumvents the need to obtain a differentiable *upper* bound on mutual information. Works by Alemi et al. [2018], Poole et al. [2019] do derive upper bounds on mutual information but not conditional mutual information. Moreover, Poole et al. [2019] require knowledge of $p(y|x)$ to arrive at their upper bound (see their Figure 1); we do not have this requirement. If $p(y|x)$ is not known, Poole et al. [2019] provide lower bounds that are a refinement of MINE [Belghazi et al., 2018].

Conditional Independence Testing: Testing CI has been well-studied as a hypothesis testing problem and is central to works on causality [Koller and Friedman, 2009, Pearl, 2009, Peters et al., 2017] and high-dimensional feature selection. Traditional methods relied on testing correlation between residuals of $Y|Z$ and $X|Z$. Works like Zhang et al. [2011], Gretton et al. [2012, 2008] extended this principle using kernel spaces; Park and Muandet [2020] do so for conditional distributions. There is a recent line of work that uses a perfect sampler from conditional distributions to accomplish independence testing [Bellot and van der Schaar, 2019, Candes et al., 2016, Berrett et al., 2020]. Recently, with the success of neural networks, so-called model-powered approaches have used strong classifiers to map the problem of CI testing to nearest neighbor estimation and classification [Sen et al., 2017]. Inspired by Sen et al. [2017], we provide a differentiable CI-*enforcing* method based on GANs. We would like to note that, generation is a different problem compared to testing when the focus is only about accepting or rejecting the null which is the conditionally independent distribution.

Fairness Criteria and Fair Classification: We mention more closely related works within the rapidly-growing literature on fair supervised learning. One line of work [Edwards and Storkey, 2016, Xie et al., 2017, Beutel et al., 2017, Zhang et al., 2018, Madras et al., 2018, Xu et al., 2018, Song et al., 2019] aims to achieve fairness through adversarial means by learning representations that remain predictive of an outcome Y but are invariant to (i.e. poorly predictive of) a sensitive attribute S. More recent works [Beutel et al., 2017, Zhang et al., 2018, Madras et al., 2018, Song et al., 2019] also address the equalized odds criterion $\hat{Y} \perp\!\!\!\perp S \,|\, Y$ and we compare to Zhang et al. [2018] herein. Similar to our work, Xu et al. [2018] use GANs to generate data $(\hat{X}, \hat{Y})$ close to the given distribution of (X, Y) while satisfying fairness conditions $\hat{X} \perp\!\!\!\perp S$ and $\hat{Y} \perp\!\!\!\perp S$. These conditions however are akin to statistical parity and are not conditional. Song et al. [2019] make use of bounds on mutual information similar to those of Alemi et al. [2018], Poole et al. [2019] cited above. However, Song et al. [2019] also do not address general conditioning, focusing on demographic parity (not conditional), equal opportunity (restricting to $Y = 1$), and equalized odds, where they exploit the binary nature of Y.

Conditional statistical parity (CSP) was introduced by Kamiran et al. [2013] and further discussed by Corbett-Davies et al. [2017]. The methods of Kamiran et al. [2013] achieve CSP by stratification and thus work best with a single discrete admissible variable A, i.e. conditioning on a scalar discrete variable. In contrast, our proposed method can handle multiple admissible variables without the exponential dependence on dimension entailed by stratification. A generalization of CSP is stated in Salimi et al. [2019] as a sufficient condition for their causal notion of *justifiable fairness*. This is a concrete example of a CI statement being used as a sufficient condition for a causal fairness definition; connections to other definitions by Kilbertus et al. [2017], Kusner et al. [2017], Nabi and Shpitser [2018], Chiappa [2019] may be possible. Salimi et al. [2019] propose algorithms based on MaxSAT and non-negative matrix factorization; the latter approach however has to enumerate all values of the conditioning variables.

2 A GENERAL CHARACTERIZATION OF CONDITIONAL INDEPENDENCE

We develop our key theoretical result for random variables defined over real domains. Let X, Y and Z be three random variables taking values in $\mathcal{X} \subseteq \mathbb{R}^{d_x}$, $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$ and $\mathcal{Z} \subseteq \mathbb{R}^{d_z}$ and following a joint distribution $P_{X,Y,Z}$. To simplify notation, we will drop the subscripts from P when it is clear that we are referring to the joint distribution. We are interested in a measure of conditional dependence of X and Y given Z. Conditional independence (CI) is written as $X \perp\!\!\!\perp Y \,|\, Z$.

For technical simplicity, we assume that P and other probability distributions to be introduced are absolutely continuous with respect to a measure ν and that their Radon-Nikodym derivatives exist, e.g. $\frac{dP}{d\nu} = p$. In particular, we focus on the case where ν is the Lebesgue measure over $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y} \times \mathbb{R}^{d_z}$ and p is therefore a density function. The same development holds for discrete distributions with p representing a probability mass function and ν the counting measure (and suitably modified proofs).

A *divergence* $D(P, Q)$ between probability distributions P and Q is usually understood to be a non-negative function $D(P, Q) \geq 0$ for all P, Q such that $D(P, Q) = 0$ if and only if $P = Q$. Following the discussion in the previous paragraph, we will consider D to be a function of the corresponding Radon-Nikodym derivatives or densities, i.e. $D(p, q)$ with $q = \frac{dQ}{d\nu}$.

Our characterization of CI involves divergences between the given distribution P of (X, Y, Z) and a distribution Q of (X, Y', Z), where the joint distribution of (X, Z) is the same as in P while $Y' \in \mathcal{Y}$ follows a conditional distribution $Q_{Y' \mid Z}$ independent of X, with conditional density function $q_{Y' \mid Z}$. Thus the marginal density of Q with respect to (Y', Z) is $q_{Y',Z} = p_Z q_{Y' \mid Z}$ and the joint density is $q = q_{X,Y',Z} = p_{X,Z} q_{Y' \mid Z}$. The choice of $q_{Y' \mid Z}$ is fairly flexible and we discuss it in Section 3.3. We use $q_{Y' \mid z}$ and similar notation to denote the conditional density of Y' for a fixed z.

To obtain our characterization of CI formally, we assume that D has the following additional properties:

Assumption 1 (Strict convexity)**.** $D(p, q)$ is a strictly convex function of either p or q.

Assumption 2 (Separability)**.** Suppose that p and q are joint densities over $\mathcal{X} \times \mathcal{Y}$ with the same marginal density with respect to X, i.e. $p = p_X p_{Y \mid X}$ and $q = p_X q_{Y \mid X}$. Then $D(p, q) = \mathbb{E}_{x \sim P_X}[D(p_{Y \mid x}, q_{Y \mid x})]$ is the expectation of the divergence between conditional distributions of Y.

Theorem 1. *Let $P_{X,Y,Z}$ and $Q_{X,Y',Z}$ be the joint distributions of (X, Y, Z) and (X, Y', Z) specified above. If divergence $D(p, q)$ is strictly convex in p (Assumption 1) and separable (Assumption 2), then*

$$D\big(p_{X,Y,Z}, q_{X,Y',Z}\big) = D\big(p_{Y,Z}, q_{Y',Z}\big) \iff X \perp\!\!\!\perp Y \mid Z.$$

All proofs can be found in the supplementary material (SM). If $D(p, q)$ is strictly convex in q instead of p as in Theorem 1, then the same result is obtained by switching the arguments of the divergence.

Corollary 1. *If $D(p, q)$ is strictly convex in q (Assumption 1) and separable (Assumption 2), then*

$$D(q_{X,Y',Z}, p_{X,Y,Z}) = D(q_{Y',Z}, p_{Y,Z}) \iff X \perp\!\!\!\perp Y \mid Z.$$

We discuss known special cases of Theorem 1 and Corollary 1 in Section 2.2.

2.1 THE DEPENDENT CASE AND A MEASURE OF DEPENDENCE

We now discuss the case in which X and Y are dependent conditioned on Z. Theorem 1 then implies that $D(p_{X,Y,Z}, q_{X,Y',Z}) \neq D(p_{Y,Z}, q_{Y',Z})$, and in fact we have $D(p_{X,Y,Z}, q_{X,Y',Z}) > D(p_{Y,Z}, q_{Y',Z})$ since the proof of Theorem 1 shows that the difference between the divergences is non-negative. Specifically, the difference is the expectation of a non-negative function

$$\xi(z) = \mathbb{E}_{x \sim P_{X \mid z}}\left[D\big(p_{Y \mid x,z}, q_{Y' \mid z}\big)\right] - D\left(\mathbb{E}_{x \sim P_{X \mid z}}\left[p_{Y \mid x,z}\right], q_{Y' \mid z}\right). \quad (1)$$

We may then interpret the magnitude of the difference $D(p_{X,Y,Z}, q_{X,Y',Z}) - D(p_{Y,Z}, q_{Y',Z})$ as a measure of conditional dependence of X and Y.

Taking this interpretation a step further, we can consider the function $\xi(z)$ as a measure of the dependence of X and Y conditioned on a particular $Z = z$. Examination of (1) shows that $\xi(z)$ is the *slack* in Jensen's inequality, i.e. the difference between the expectation of a convex function of $p_{Y \mid x,z}$ and the same convex function evaluated at the expected value of $p_{Y \mid x,z}$, which is $p_{Y \mid z}$. Qualitatively speaking, the more that $p_{Y \mid x,z}$ varies with (i.e. depends on) x, the greater the slack $\xi(z)$ is expected to be. If $p_{Y \mid x,z}$ does not vary with x (almost surely), then $\xi(z) = 0$.

With additional assumptions, it is possible to relate $\xi(z)$ to a measure of variation with x based on $\mathcal{L}_2$ distance between $p_{Y \mid x,z}$ and $p_{Y \mid z}$. The derivation is in the SM.

Proposition 1. *Assume that $D(p, q)$ is differentiable and strongly convex in p with parameter m, and that $p_{Y \mid z}$, $p_{Y \mid x,z}$ for all x such that $p_{X \mid z}(x \mid z) > 0$, and $\nabla_p D(p, q_{Y' \mid z})|_{p = p_{Y \mid z}}$ all belong to the space of square-integrable functions $\mathcal{L}_2(\mathcal{Y})$. Then*

$$\xi(z) \geq \frac{m}{2} \mathbb{E}_{x \sim P_{X \mid z}}\left[\left\| p_{Y \mid x,z} - p_{Y \mid z} \right\|_{\mathcal{L}_2}^2\right].$$

2.2 DIVERGENCES SATISFYING ASSUMPTIONS

We show that many well-known divergences satisfy Assumptions 1 and 2, and therefore Theorem 1 and/or Corollary 1 apply to them.

f-divergences Given two distributions P and Q with densities $p(x)$, $q(x)$ such that P is absolutely continuous with respect to Q, and a convex function $f : \mathbb{R}_+ \mapsto \mathbb{R}$ such that $f(1) = 0$, the f-divergence between P and Q is defined as

$$D_f(p, q) = \mathbb{E}_Q\left[f\left(\frac{p(X)}{q(X)}\right)\right]. \quad (2)$$

Due to the fact that $p(x)$ enters into (2) only through the ratio $p(x)/q(x)$, all f-divergences satisfy the separability property (Assumption 2) as verified in the SM. Assumption 1 can be satisfied if the function f is strictly convex.

Proposition 2. *If $f : \mathbb{R}_+ \mapsto \mathbb{R}$ is strictly convex, then $D_f(p, q)$ is a strictly convex function of p for all p that are absolutely continuous with respect to q.*

Remark: In the case of f-divergences, the above indicates that absolute continuity of p with respect to q is needed for Theorem 1 to hold. For bounded Y, we can take q to be the uniform distribution.

It follows that many common f-divergences satisfy Assumptions 1 and 2: KL divergence ($f(t) = t \log t$ or $f(t) = -\log t$), χ^2 divergence ($f(t) = t^2 - 1$), squared Hellinger distance ($f(t) = 2(1 - \sqrt{t})$), but not total variation distance.

Kullback-Leibler divergence In the case of KL divergence, $\mathrm{KL}(p \,\|\, q) = \mathbb{E}_P \left[\log\left(p(X)/q(X)\right)\right]$, Theorem 1 reduces to the well-known condition of conditional mutual information being zero.

Corollary 2. *If D is the Kullback-Leibler divergence, then Theorem 1 reduces to $I(X; Y \mid Z) = 0 \iff X \perp\!\!\!\perp Y \mid Z$, where $I(X; Y \mid Z)$ is the conditional mutual information.*

In this case, the auxiliary variable Y' drops out of the identity.

f-divergences without conditioning If Z is constant, then we may drop the conditioning on Z and drop Z from all distributions. Then the identity in Theorem 1 becomes $D(p_{X,Y}, q_{X,Y'}) \geq D(p_Y, q_{Y'})$, where again the inequality is shown in the proof. If we also let Y' have a general conditional distribution $q_{Y' \mid X}$, then this coincides with the "conditioning increases f-divergence" property of Polyanskiy and Wu [2019, Thm. 6.1].Theorem 1 is more general because 1) we do condition on arbitrary Z, as required in our application, and 2) we do not restrict ourselves to f-divergences, instead identifying general conditions on the divergence (Assumptions 1 and 2) for the theorem to hold.

Jensen-Shannon divergence The case of Jensen-Shannon (JS) divergence is of particular interest in this paper because it forms the basis for the architecture in Section 3. We use the following definition of JS divergence between distributions P and Q with densities p and q:

$$\mathrm{JS}(p \,\|\, q) = \frac{1}{2}\mathrm{KL}\left(p \,\middle\|\, \frac{p+q}{2}\right) + \frac{1}{2}\mathrm{KL}\left(q \,\middle\|\, \frac{p+q}{2}\right). \tag{3}$$

The JS divergence is also a (perhaps less well-known) f-divergence with $f(t) = \frac{t}{2}\log\left(\frac{2t}{1+t}\right) + \frac{1}{2}\log\left(\frac{2}{1+t}\right)$. The first term in $f(t)$ has been noted e.g. by Lin [1991]. Since $f''(t) = 1/(2t(1+t)) > 0$ for $t > 0$, $f(t)$ is strictly convex. Hence the JS divergence satisfies both Assumptions 1 and 2.

Bregman divergences Let F be a strictly convex and differentiable function mapping probability distributions to the reals. The function F defines a Bregman divergence through

$$D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle, \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes an inner product. Bregman divergences thus satisfy Assumption 1 by virtue of (4) and the strict convexity of F. Besides KL divergence (and its generalizations), a Bregman divergence that also satisfies Assumption 2 is Itakura-Saito distance, due to the fact that it depends on (p, q) only through their ratio, similar to f-divergences [Banerjee et al., 2005].

3 CONDITIONALLY INDEPENDENT DATA GENERATION

In the remainder of the paper, we consider the problem of generating data from a distribution that satisfies a desired CI statement while remaining close to a given data distribution. We now use X, Y, Z, W to denote random variables that are distributed according to the given distribution, with density $p(x, y, z, w)$.

Our goal is to generate samples $(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w})_i$ from the same domain $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} \times \mathcal{W}$ and following a distribution $\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w})$ that is close to the input distribution $p(x, y, z, w)$ in JS divergence, while ensuring that $\tilde{X}$ is conditionally independent of $\tilde{Y}$ given $\tilde{Z}$. The optimization is stated as

$$\min \; \mathrm{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w}) \,\|\, p(x, y, z, w)\big) \quad \text{s.t.} \quad \tilde{X} \perp\!\!\!\perp \tilde{Y} \mid \tilde{Z}. \tag{5}$$

We leverage the results of Section 2 by assuming that we have a sampler for $Y' \sim q(y' \mid z_f)$ such that $\tilde{p}(y \mid \tilde{z})$ is positive only where $q(y \mid \tilde{z}) > 0$ a.s. This ensures that the joint densities $\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z})$ and $q(\tilde{x}, y', \tilde{z}) = \tilde{p}(\tilde{x}, \tilde{z}) q(y' \mid \tilde{z})$ satisfy the absolute continuity assumption in Proposition 2, which in turn ensures that Assumption 1 and Theorem 1 hold. We then proceed to use the Jensen-Shannon version of Theorem 1 and the dependence measure that it defines to relax (5) as follows:

$$\begin{aligned} \min \quad & \mathrm{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w}) \,\|\, p(x, y, z, w)\big) \\ \text{s.t.} \quad & \mathrm{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}) \,\|\, q(\tilde{x}, y', \tilde{z})\big) - \mathrm{JS}\big(\tilde{p}(\tilde{y}, \tilde{z}) \,\|\, q(y', \tilde{z})\big) \leq \delta. \end{aligned} \tag{6}$$

The choice of JS divergence allows us to exploit the capabilities of GANs, as described in Section 3.1.

Remark: If $W = \emptyset$, then as discussed in the introduction, (5) could be addressed by faithfully generating $\tilde{Y}$ following $p(y|z)$. Conditional generation however becomes more difficult in high dimensions. When W is non-empty, there is an additional trade-off between CI constraint imposition and closeness between $\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}, \tilde{w})$ and $p(x, y, z, w)$. For example, if the generated data is used to learn a predictor for $\tilde{Y}$, one may not want to hurt accuracy too much by completely ignoring W just to satisfy conditional independence amongst $\tilde{Y}, \tilde{X}$ and $\tilde{Z}$.

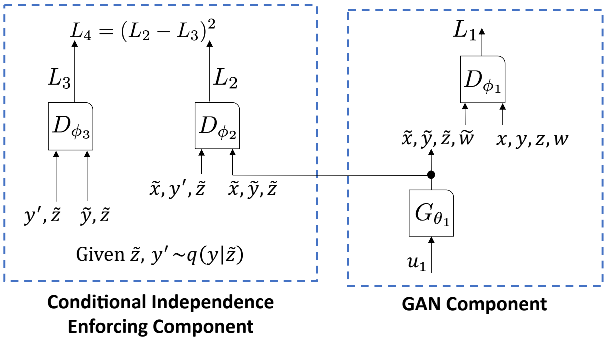

Figure 1: Proposed architecture to enforce conditional independence

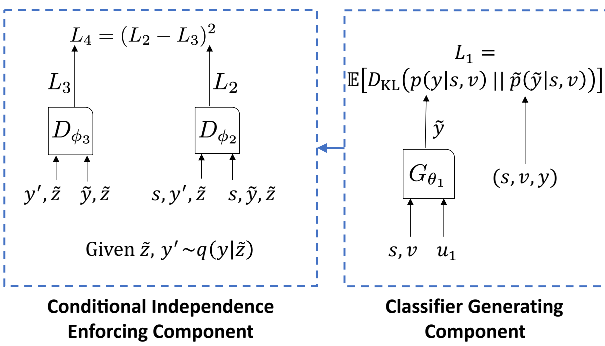

Figure 2: Simplified architecture to enforce fairness

3.1 GENERAL GAN ARCHITECTURE AND ALGORITHM

We propose using the general GAN architecture provided in Figure 1 for generating samples from a distribution $\tilde{p}$ that aims to solve (6). The architecture involves three discriminators D_{ϕ_i}, $i \in \{1, 2, 3\}$, one generator G_{θ_1}, and a sampler which samples $Y' \sim q(y'|\tilde{z})$. The generator and the first discriminator D_{ϕ_1} constitute a typical GAN which attempts to bring the generated distribution closer to the original one. Discriminators D_{ϕ_2}, D_{ϕ_3} together with loss L_4 comprise the CI-enforcing component. The two discriminators compute tight variational lower bounds L_2 and L_3 on the two JS divergences in the constraint in (6). Loss L_4 then encourages the squared difference $(L_2 - L_3)^2$ to be small; other functions of the difference are possible.

The loss functions of the three discriminators are standard GAN losses that approximate the JS divergences between the distributions whose samples are given as input [Nowozin et al., 2016]. Specifically,

$$\begin{aligned} L_1 &= \mathbb{E}_{u_1}[\log(1 - D_{\phi_1}(G_{\theta_1}(u_1)))] \\ &\quad + \mathbb{E}_{(x,y,z,w)\sim p(x,y,z,w)}[\log D_{\phi_1}(x, y, z, w)] \\ L_2 &= \mathbb{E}_{u_1}[\log(1 - D_{\phi_2}(G_{\theta_1}(u_1)))] \\ &\quad + \mathbb{E}_{(\tilde{x},y',\tilde{z})}[\log D_{\phi_2}(\tilde{x}, y', \tilde{z})] \\ L_3 &= \mathbb{E}_{\tilde{y},\tilde{z}}[\log(1 - D_{\phi_3}(\tilde{y}, \tilde{z}))] + \mathbb{E}_{(y',\tilde{z})}[\log D_{\phi_3}(y', \tilde{z})]. \end{aligned} \tag{7}$$

Here, $D_\omega(x) = \frac{1}{1+e^{-V_w(x)}}$ is the sigmoid function acting on the logit output $V_w(x)$ of a deep neural network parameterized by ω.

The training of the weights in the architecture proceeds as specified in Algorithm 1. Below we describe the two alternating steps that correspond to lines 4–5 and 6–7 in Algorithm 1.

Training Discriminators: Keeping θ_1 fixed, the three discriminators maximize their corresponding losses L_1, L_2, L_3 with respect to their parameters ϕ_1, ϕ_2 and ϕ_3, thus approximating the JS divergences between the input distributions to the discriminators.

Training Generator: Keeping the discriminator parameters ϕ_1, ϕ_2, ϕ_3 fixed, the generator is trained to optimize the combination of two losses, one that enforces similarity between the given and generated distributions (L_1), and one that ensures the desired CI (L_4). The generator objective is

$$\min \gamma L_4 + L_1, \tag{8}$$

where γ is used as a trade-off parameter. Note that the generator minimizes only the (squared) difference between losses $(L_2 - L_3)^2$ and not L_2, L_3 themselves.

3.2 THEORETICAL RESULTS

In this subsection, our interest is in showing that if discriminators approximate the divergences well, then large conditional dependence necessarily implies a large value for our metric and conditional independence would imply small value for our metric. The following lemma asserts that the losses L_2, L_3 provide variational lower bounds on their respective JS divergences. The proof follows from Sections 2.1 and 2.4 in Nowozin et al. [2016].

Lemma 1. *For any θ_1, ϕ_2, and ϕ_3 we have:*

$$\begin{aligned} L_2 &\le 2\mathrm{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}) \,\|\, q(\tilde{x}, y', \tilde{z})\big) - \log 4, \\ L_3 &\le 2\mathrm{JS}\big(\tilde{p}(\tilde{y}, \tilde{z}) \,\|\, q(y', \tilde{z})\big) - \log 4. \end{aligned}$$

The next result simply makes precise the fact that if L_2 and L_3 are close approximations to the JS divergences, then their difference reflects the dependence measure in the constraint in (6). It is a consequence of Lemma 1 and Theorem 1.

Algorithm 1 Conditionally Independent Data Generation

.

Input: Dataset: $D \sim p(x, y, z, w)$; **Iterations:** T_1, T_2, E; **Stepsizes:** η_1, η_2; **Sampler:** Given $\tilde{z}$ samples $y' \sim q(y'|\tilde{z})$.
Initialize: Set parameters $\phi_1, \phi_2, \phi_3, \theta_1$ randomly, and iteration counter $e = 1$.
for $e = 1, \dots, E$ **do**
  for $t_1 = 1, \dots, T_1$ **do**
    $(\phi_1, \phi_2, \phi_3) \leftarrow \texttt{GRADIENT DESCENT}(-L_3 - L_2 - L_1, \eta_1, (\phi_1, \phi_2, \phi_3))$ ▷ Train Discriminators
  for $t_2 = 1, \dots, T_2$ **do**
    $\theta_1 \leftarrow \texttt{GRADIENT DESCENT}(L_1 + \gamma L_4, \eta_2, \theta_1)$ ▷ Train Generator
Output: Generator G_{θ_1}.

Proposition 3. *For a given θ_1, suppose that there exist ϕ_2^* and ϕ_3^* that provide ϵ-approximations to their respective JS divergences:*

$$L_2 \geq 2\text{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}) \,\|\, q(\tilde{x}, y', \tilde{z})\big) - \log 4 - \epsilon,$$
$$L_3 \geq 2\text{JS}\big(\tilde{p}(\tilde{y}, \tilde{z}) \,\|\, q(y', \tilde{z})\big) - \log 4 - \epsilon.$$

Then

$$L_2 - L_3 \geq 2\big(\text{JS}\big(\tilde{p}(\tilde{x}, \tilde{y}, \tilde{z}) \,\|\, q(\tilde{x}, y', \tilde{z})\big) - \text{JS}\big(\tilde{p}(\tilde{y}, \tilde{z}) \,\|\, q(y', \tilde{z})\big)\big) - \epsilon,$$

and if conditional independence holds, i.e. $\tilde{Y} \perp\!\!\!\perp \tilde{X} \mid \tilde{Z}$, we also have $L_2 - L_3 \leq \epsilon$.

In particular, minimizing $L_4 = (L_2 - L_3)^2$ brings $\tilde{X}, \tilde{Y}, \tilde{Z}$ closer to conditional independence, provided that L_2 and L_3 approximate the JS divergences well.

Based on the development in Section 2, the above "difference of divergences" dependence measure assumes absolute continuity of p with respect to q. In the theorem below however, we use the particular properties of JS divergence to show that even if absolute continuity is not satisfied, $L_2 - L_3$ is still bounded from below by a different dependence measure.

Theorem 2. *For a given θ_1, suppose that ϕ_2^* and ϕ_3^* satisfy the assumptions of Proposition 3. For some constants $\eta_1 > 0$, $\eta_2 > 0$, $\delta > 0$, $\gamma > 0$, suppose the distribution q is such that the event $\mathcal{B}(\eta_1, \eta_2, \delta) = \{(x, y, z) : \tilde{p}(y|z) > \eta_1,\ q(y|z) \geq \eta_2,\ |\log \frac{\tilde{p}(y|x,z)}{\tilde{p}(y|z)}| > \delta\}$ has probability $\Pr_{\tilde{p}(x,y,z)}(\mathcal{B}(\eta_1, \eta_2, \delta)) > \gamma$. Then*

$$L_2 - L_3 \geq 2\gamma \frac{\eta_1^2 (1 - e^{-\delta})^2}{(1 + 1/\eta_2)^4} - \epsilon.$$

Remark: Theorem 2 quantifies dependence in terms of the probability γ of regions where the log-ratio $\log \frac{\tilde{p}(y|x,z)}{\tilde{p}(y|z)}$ is large and both $\tilde{p}(y|z)$ and $q(y|z)$ have non-zero probability mass. In particular, it stipulates that q should have probability mass in regions where $\tilde{p}$ has mass *and* conditional dependence is high. This is weaker than absolute continuity of $\tilde{p}$ with respect to q. Note also that the lower bound on $L_2 - L_3$ is increasing in all parameters $\eta_1, \eta_2, \delta, \gamma$.

3.3 CHOICE OF THE SAMPLING DISTRIBUTION q

If $\mathcal{Y}$, the domain of Y, $\tilde{Y}$, and Y', is bounded or discrete with finite cardinality, then it suffices to choose the sampling distribution $q(y' \mid \tilde{z})$ to be uniform over the support. This ensures that $q(y \mid \tilde{z})$ covers the support of $\tilde{p}(y \mid \tilde{z})$ completely. It also resolves any support issues in estimating JS divergence by discriminators D_{ϕ_2} and D_{ϕ_3}, so that losses L_2 and L_3 will not diverge to infinity even if discriminator training is run for longer. In fairness applications in Section 4, Y can be taken to be a scalar outcome variable, i.e. $d_y = 1$, and in classification settings it has finite cardinality. We therefore adopt uniform sampling in the experiments in Section 4.2.

4 APPLICATIONS TO ENFORCING FAIRNESS CRITERIA

We discuss an application of the framework of Section 3 to fairness in machine learning, and specifically to enforcing two fairness measures that involve conditioning. The first condition is *conditional statistical parity* (CSP) [Kamiran et al., 2013] where we wish to make outcomes independent of protected attributes conditioned on *admissible variables* A, i.e. $\tilde{Y} \perp\!\!\!\perp S \mid A$. The well-known Berkeley admissions case [Bickel et al., 1975] makes clear the importance of CSP, where the bias in admissions ($\tilde{Y}$) against female applicants (S is gender) changed patterns when conditioned on departments (A). In the CSP case, the advantage of the proposed CI generation method is that it handles multiple admissible variables (possibly continuous) while avoiding enumeration of all their values. The second fairness criterion is *equalized odds* (EO) [Hardt et al., 2016], a well-known measure used in fair binary classification. It requires equal rates of false positives and false negatives between groups defined by *protected attributes* S. Denoting the predicted and true labels by $\tilde{Y}$ and Y, this corresponds to $\tilde{Y} \perp\!\!\!\perp S \mid Y$. As mentioned in the introduction, there are many existing methods for enforcing EO, and our consideration of EO can be seen more as a proof of concept that CI data generation works in a known setting.

Table 1: Differences in accuracy and differences in maximum conditional statistical disparity (MCSD) with respect to $\gamma = 0$. Protected attribute is gender. Admissible attributes are years of education (top) and both years of education and hours per week (bottom). Standard errors in parentheses.

γ	ΔAcc. (%)	ΔMCSD$_{\text{edu}}$ (%)
0.01	0.2 (0.3)	1.1 (1.9)
0.1	0.3 (0.3)	-0.6 (1.6)
1.0	-0.2 (0.4)	-1.8 (2.0)
10	-0.9 (0.4)	-7.1 (2.1)
50	-2.9 (0.4)	-17.3 (2.4)
100	-2.3 (0.4)	-16.9 (2.5)
1000	-3.0 (0.3)	-23.7 (1.7)

γ	ΔAcc. (%)	ΔMCSD$_{\text{edu}}$ (%)	ΔMCSD$_{\text{hrs}}$ (%)
0.01	0.0 (0.3)	-0.6 (1.8)	3.2 (2.9)
0.1	-0.2 (0.3)	-0.9 (2.1)	-2.6 (2.7)
1.0	-0.5 (0.3)	-1.0 (2.1)	-4.7 (2.8)
10	-0.2 (0.4)	-6.2 (2.2)	-12.7 (2.5)
50	-3.0 (0.3)	-18.3 (2.6)	-25.0 (2.1)
100	-2.8 (0.4)	-20.6 (2.6)	-23.8 (2.2)
1000	-5.5 (1.9)	-16.0 (2.7)	-18.7 (2.6)

Table 2: Changes in accuracy and equalized odds difference (EOD) for the proposed CI method (with respect to $\gamma = 0$) and adversarial debiasing (AD) [Zhang et al., 2018] (with respect to $\lambda_a = 0$). Protected attribute is gender.

CI γ	ΔAcc. (%)	ΔEOD (%)
0.01	0.0 (0.3)	-1.2 (0.6)
0.1	0.1 (0.3)	-0.2 (0.6)
1.0	0.2 (0.3)	-1.2 (0.7)
10	-1.4 (0.4)	-3.8 (0.6)
30	-1.3 (0.3)	-3.1 (0.6)
50	-1.3 (0.3)	-4.0 (0.6)
100	-2.0 (0.3)	-4.6 (0.6)
200	-2.8 (0.4)	-3.5 (1.0)
300	-2.8 (0.3)	-5.0 (0.5)

AD λ_a	ΔAcc. (%)	ΔEOD (%)
0.01	-0.1 (0.2)	-0.5 (0.3)
0.1	-1.1 (0.3)	1.1 (0.4)
1.0	-1.8 (0.2)	10.6 (1.4)
10	-7.1 (0.4)	16.8 (5.7)

4.1 ARCHITECTURE

In Figure 2, we specialize the generic architecture proposed in Figure 1 to promote CSP and EO. The sensitive attributes S play the role of $\tilde{X}$. In the CSP case, the conditioning variable $\tilde{Z} = A$, the admissible variables, whereas in the EO case, $\tilde{Z}$ maps to Y, the true label. The symbol V represents all predictor variables other than the sensitive attributes, including admissible variables A and other variables W.

The major difference in Figure 2 is that only the binary $\tilde{Y}$ is generated while $\tilde{X} = S$ and $\tilde{Z} = A$ or $\tilde{Z} = Y$ come from the original data. Hence, this is a simpler special case. As a consequence, the generator G_{θ_1} reduces to a classifier that takes the feature set (S, V) and outputs a predicted label $\tilde{Y}$ such that the cross-entropy loss between the ground truth and predicted label distributions is small. This cross-entropy loss takes on the role of discriminator D_{ϕ_1} in Figure 1. The other components on the left side remain the same.

4.2 EXPERIMENTS

We demonstrate the utility of the architecture in Figure 2 for fair classification on the *Adult Census Income* [Kohavi, 1996] dataset. The target variable is whether a person's annual income exceeds $50,000$ USD. We consider gender/sex and race as the protected attributes. For the CSP experiments, we consider years of education and hours worked per week as admissible attributes since these are well-accepted as legitimate determinants of income. We use the dataset's fixed train/test split and report results on the test set. Additionally, we held out 30% of the training samples as the validation set. The SM contains further details on data preprocessing, the architecture and optimization. We report the mean and standard error over 25 runs for the metrics.

Conditional Statistical Parity Results We implemented the architecture in Figure 2 for CSP. Here we take gender as the protected attribute and evaluate *maximum conditional statistical disparity* (MCSD) by first computing the difference between predicted positive rates for females and males, conditioned on each value of the admissible variable, and then taking the maximum absolute difference. In the unpenalized case, our CI method with $\gamma = 0$ in (8) achieves an accuracy of $(82.6 \pm 0.2)\%$. Our main findings are illustrated in Table 1, which shows *differences* in accuracy and differences in MCSD (denoted by Δ) with respect to the $\gamma = 0$ values as γ is increased. With years of education as the admissible variable (corresponding to Table 1, top), the baseline MCSD for $\gamma = 0$ is $(38.2 \pm 1.4)\%$, whereas with both education and work hours per week as admissible variables (Table 1 bottom), the baseline MCSD is $(37.5 \pm 1.1)\%$ for education (averaging out hours/week) and $(34.8 \pm 1.9)\%$ for hours/week (averaging out education). We see that increasing γ reduces MCSD without a substantial reduction in accuracy.

Equalized Odds Results For EO, we compare with the adversarial debiasing (AD) [Zhang et al., 2018] algorithm as a point of reference. AD was chosen because it is also a GAN-like solution, developed specifically for fairness. Adherence to EO is measured by the *average absolute equalized odds difference* (EOD), which is the average of the absolute differences in false positive rate (FPR) and negative rate

(FNR) between two protected groups. In the unpenalized case, our CI method with $\gamma = 0$ in (8) achieves an accuracy of $(82.6 \pm 0.2)\%$ and an EOD of $(6.0 \pm 0.5)\%$. AD with parameter $\lambda_a = 0$ achieves $(85.2 \pm 0.1)\%$ accuracy and $(4.2 \pm 0.2)\%$ EOD. These starting metrics are different for CI and AD because of implementation differences that are unfortunately hard to reconcile. Similar to the CSP case, Table 2 shows *changes* in accuracy and EOD with respect to the $\gamma = 0$ or $\lambda_a = 0$ values as γ and λ_a are increased. For CI, increasing γ enforces EO more strictly as expected, while accuracy decreases modestly. For AD however, the EOD decreases only slightly before results deteriorate, with a large decrease in accuracy and unexpected increase in EOD. We did not increase λ_a further for this reason.

We also consider multiple protected attributes, namely sex and race together. While AD can in principle be applied to this setting by encoding sex and race as a single 4-category variable, it requires changing the discriminator loss to multi-class and we have been unable to tune it to obtain reasonable results. In contrast, CI naturally handles multiple protected attributes. For $\gamma = 0$, the EOD between sexes is $(4.8 \pm 0.4)\%$ after averaging out race, and $(6.2 \pm 0.4)\%$ between races after averaging out sex. For $\gamma = 10$, these numbers decrease to $(2.8 \pm 0.2)\%$ and $(4.7 \pm 0.7)\%$ respectively, thus improving EO with respect to both attributes, while accuracy is unchanged.

We note as a limitation that GANs are known to exhibit instability and difficulty in training, and the proposed architecture does inherit these issues.

5 CONCLUSION

We have addressed the problem of *enforcing* conditional independence (CI) on a data-generating distribution, as a complement to the large literature on testing data distributions for CIs. Underpinning the work is a flexible characterization of CI in the form of an identity that holds for a wide class of divergences. This identity formed the basis for a differentiable GAN-based architecture for generating data to balance adherence to a desired CI with proximity to a given data distribution. We demonstrated an application to enforcing the fairness criteria of equalized odds and conditional statistical parity.

One specific item for future work concerns the sampling distribution $q(y' \mid \tilde{z})$: while we have found a uniform distribution to be sufficient in our experiments, it would be interesting to explore alternatives that cover the support of $\tilde{p}(\tilde{y} \mid \tilde{z})$ and perhaps attempt to approximate it. More broadly, the proposed CI-enforcing GAN exploits the Jensen-Shannon version of the divergence identity, and fairness is only one application of conditionally independent data generation. Regarding the first point, similar architectures might be explored in future work, for example for other f-divergences, building on f-GANs [Nowozin et al., 2016]. Regarding other applications, one that would be interesting to explore is invariant prediction [Peters et al., 2016, Arjovsky et al., 2019], which can be stated as a CI condition: predictions should be independent of the environment conditional on a transformation of the data. It may also be possible to turn the proposed difference in divergences measure into a CI *testing* principle; this would require characterizing its distribution under the null hypothesis.

Acknowledgements

The authors thank the anonymous reviewers of this and previous version of the paper. Kartik Ahuja acknowledges the support provided by IVADO postdoctoral fellowship funding program.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 159–168, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/alemi18a.html.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, January 2005.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. *arXiv preprint arXiv:1907.04068*, 2019.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, pages 1–5, August 2017.

Peter J Bickel, Eugene A Hammel, and J William O'Connell. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

Emmanuel Candes, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*, 2016.

Silvia Chiappa. Path-specific counterfactual fairness. In *AAAI Conference on Artificial Intelligence (AAAI)*, January 2019.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14, May 2016.

W. Gao, S. Oh, and P. Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*, pages 1–1, 2018. ISSN 0018-9448. doi: 10.1109/TIT.2018.2807481.

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Breaking the bandwidth barrier: Geometrical adaptive entropy estimation. In *Advances in Neural Information Processing Systems*, pages 2460–2468, 2016.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 585–592. Curran Associates, Inc., 2008.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13(1):723–773, March 2012. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=2503308.2188410`.

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

Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, June 2013. doi: 10.1007/s10115-012-0584-8. URL `https://doi.org/10.1007/s10115-012-0584-8`.

Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 656–666, December 2017. URL `http://dl.acm.org/citation.cfm?id=3294771.3294834`.

Ron Kohavi. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996. URL `https://archive.ics.uci.edu/ml/datasets/adult`.

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009. ISBN 0262013193, 9780262013192.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4066–4076. December 2017. URL `http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf`.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37 (1):145–151, January 1991.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3384–3393, July 2018. URL `http://proceedings.mlr.press/v80/madras18a.html`.

Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 1931–1940. NIH Public Access, February 2018.

Morteza Noshad, Yu Zeng, and Alfred O Hero. Scalable mutual information estimation using dependence graphs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2962–2966. IEEE, 2019.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279, 2016.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21247–21259, 2020. URL `https://proceedings.neurips.cc/paper/2020/file/`

f340f1b1f65b6df5b5e3f94d95b11daf-Paper.pdf.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 052189560X, 9780521895606.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, pages 947–1012, 2016.

Barnabás Póczos and Jeff G Schneider. Nonparametric estimation of conditional information and divergences. In *AISTATS*, pages 914–923, 2012.

Yury Polyanskiy and Yihong Wu. Lecture notes on information theory, 15 May 2019. URL http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pages 5171–5180, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/poole19a.html.

Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Capuchin: Causal databbse repair for algorithmic fairness, February 2019. arXiv e-print https://arxiv.org/abs/1902.08283.

Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, pages 2951–2961, 2017.

Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 89, pages 2164–2173, 16–18 Apr 2019. URL http://proceedings.mlr.press/v89/song19a.html.

Ioannis Tsamardinos, Constantin F Aliferis, Alexander R Statnikov, and Er Statnikov. Algorithms for large scale markov blanket discovery. In *FLAIRS conference*, volume 2, pages 376–380, 2003.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 585–596. December 2017.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. FairGAN: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (Big Data)*, pages 570–575, December 2018. doi: 10.1109/BigData.2018.8622525.

Karren Yang, Abigail Katcoff, and Caroline Uhler. Characterizing and learning equivalence classes of causal dags under interventions. In *International Conference on Machine Learning*, pages 5541–5550. PMLR, 2018.

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 (WWW)*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on Artifical Intelligence, Ethics, and Society (AIES)*, pages 335–340, February 2018. doi: 10.1145/3278721.3278779. URL http://doi.acm.org/10.1145/3278721.3278779.

Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pages 804–813, Arlington, Virginia, United States, 2011. AUAI Press. ISBN 978-0-9749039-7-2. URL http://dl.acm.org/citation.cfm?id=3020548.3020641.