Learning Smooth and Fair Representations

Xavier Gitiaux **Huzefa Rangwala**
George Mason University

Abstract

This paper explores the statistical properties of fair representation learning, a preprocessing method that preemptively removes the correlations between features and sensitive attributes by mapping features to a fair representation space. The demographic parity of a representation can be certified from a finite sample if and only if the chi-squared mutual information between features and representations is finite for all features distributions. Empirically, we find that smoothing representations provides generalization guarantees of fairness certificates, which improves upon existing fair representation learning approaches. On four datasets we simulate many downstream users and show that our approach, AGWN, is the only one that generates representations whose fairness properties are robust to many downstream users.

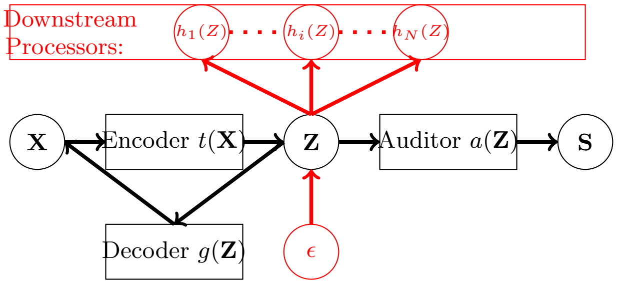

Figure 1: Fair representation learning. Variables are: features $\mathbf{X}$; sensitive attribute $\mathbf{S}$; representation $\mathbf{Z}$. Fair representation protocol includes an encoder t that maps X to its representation Z; a decoder g that reconstructs X from Z; and, an auditor a that measures the statistical dependence between Z and S. The contribution of this paper is to introduce an additive Gaussian white noise (AGWN) channel ϵ so that fairness guarantees can be established for all data processors h using Z.

1 Introduction

Organizations dealing with data are increasingly accountable for the collection, use and disposal of the data, including the responsibility of discriminatory use on the basis of sensitive attributes (e.g. racial or ethnic origin, sexual orientation or political beliefs). However, these organizations, henceafter data controllers, cannot always anticipate and control how downstream applications, henceafter data processors, will process the data. This is problematic since a growing body of evidence has raised concerns about the fairness of machine learning outcomes across a wide range of applications, including judicial decisions (ProPublica (2016)), face recognition (Buolamwini and Gebru (2018)), degree completion (Gardner et al. (2019)) or medical treatment (Pfohl et al. (2019)).

One promising avenue is to limit the data access to its fair representations (e.g Madras et al. (2018), Creager et al. (2019), Edwards and Storkey (2015), Pfohl et al. (2019) or Zemel et al. (2013)). Fair representation learning seeks to map the original data distribution into a distribution that retains the information contained in the original data, while being statistically independent of sensitive attributes (see Figure 1). However, current fair representation learning approaches provide fairness guarantees only against *some* pre-specified data processors (Chouldechova and Roth (2018)). *This paper explores conditions on the encoder to generate representation distributions with fairness guarantees that hold for any data processor.*

We show that for fairness guarantees derived from finite samples to generalize to all downstream data processors, it is necessary that the χ^2 mutual information between feature and representation is finite. Moreover, we prove that a finite χ^2 mutual information between

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

feature and representation is a sufficient condition on representation mappings to guarantee a good approximate rate ($O(n^{-1/2})$) of empirical certificates.

In practice, it is challenging to establish that the χ^2 mutual information is finite without knowing the distribution over $\mathcal{X}$. However, we show that an additive Gaussian white noise (AGWN) channel placed after any representation mapping (see Figure 1) will bound the χ^2 mutual information once the representations have passed through the channel. The channel smoothes the representation distribution by transforming it into a mixture of Gaussian distributions that can be estimated by Monte Carlo integration (Goldfeld et al. (2019)). Therefore, a plug-in fairness auditor that relies on estimating the class conditional density functions over the representation space achieves a convergence rate of $O(n^{-1/2})$ and thus delivers meaningful empirical certificates of fairness.

We empirically find on various synthetic and fair learning benchmark datasets that an AGWN channel in fair representation learning is sufficient for empirical certificates to upper bound the demographic disparity of multiple downstream users that attempts to predict sensitive attributes from samples of the representation distribution. An AGWN channel improves upon existing approaches in adversarial fair representation learning whose fairness guarantees do not extend beyond a set of specific downstream users. Moreover, we find that obtaining good approximation rates for empirical certificates does not come at the cost of significantly degrading the accuracy-fairness trade-off of downstream predictive tasks.

Related work. A growing literature explores the potential adverse implications that machine learning algorithms might have on protected demographic groups (e.g individuals self-identified as Female or African-American) (Chouldechova and Roth (2018) for a review). Many contributions seek to define fairness criteria either at the group or individual level (Dwork et al. (2012)) and then, impose a fairness penalty into their classification algorithm (e.g. Agarwal et al. (2018), Kim et al. (2018), Kearns et al. (2018)) or audit for a specific criteria (e.g Feldman et al. (2015), Gitiaux and Rangwala (2019)). In this paper, we side-step the important discussion on what fairness criteria to choose from (Kleinberg et al. (2016)), but investigate whether a data can be transformed so that any future use will meet a pre-specified criteria. Our results focus on demographic parity (Dwork et al. (2012)), but can be readily extended to many other group level criteria, including equalized odds and equal opportunity (Hardt et al. (2016)).

Existing pre-processing methods to mitigate unfair data use include sampling and reweighting (e.g. Calders and Žliobaitė (2013), Gordaliza et al. (2019)), optimization procedures to learn a data transformation that both preserve utility and limit discrimination (e.g. Calmon et al. (2017)), and representation learning (e.g. Zemel et al. (2013)). Representation learning seeks to encode the data while removing correlations between features and sensitive attributes. A data encoder generates a representation of the data and fools a neural network that attempts to predict sensitive attributes from samples of the representation distribution (e.g. Edwards and Storkey (2015), Madras et al. (2018), Zhang et al. (2018) or Xu et al. (2018)). An alternative approach is to disentangle sensitive attributes from features by passing the data through an information bottleneck (Louizos et al. (2015) or Creager et al. (2019)).

Our contribution to fair adversarial learning is to explore conditions so that the learned representation offers fairness guarantees against downstream processors that do not necessarily belong to the same class as the adversary used during the training of the encoder. Madras et al. (2018) and Oneto et al. (2019) explore empirically whether representations that achieve demographic parity for a specific downstream task generalize to new tasks in terms of accuracy and fairness. We extend their work by showing theoretically and empirically that introducing an AGWN channel in fair representation learning offers generalization guarantees to all future tasks. Moreover, introducing an AGWN channel avoids the need for an adversarial auditor, since it allows approximating the empirical fairness certificate with a differentiable loss that can be computed by Monte Carlo sampling.

Similar to our approach, the differential privacy literature relies on noise injection to guarantee that two neighboring datasets are indistinguishable (Dwork et al. (2014)). However, in the context of differential privacy, indistinguishability is only obtained by adding Gaussian/Laplacian noise. In our fairness context, for a finite sample, statistical hiding comes from learning representations subject to a demographic parity constraint; the injection of Gaussian noise is only a means to generalize the statistical hiding property to the infinite sample regime.

2 Certifying Fair Representations

2.1 Background

Consider a data controller who wants to release samples from a distribution μ over $\mathcal{X} \times \mathcal{S}$ with features in $\mathcal{X} \subset [0, 1]^D$ and sensitive attributes in $\mathcal{S}$. Although our setup can be extended to richer spaces of sensitive attributes, we focus here on binary sensitive attributes and assume that $\mathcal{S} = \{0, 1\}$.

A transformation t that maps the features space $\mathcal{X}$ into a representations space $\mathcal{Z} \subset \mathbb{R}^d$ induces a distribution μ_t over $\mathcal{Z} \times \{0, 1\}$: $\mu_t(A) = \mu\left(\{x \in \mathcal{X} | t(x) \in A\}\right)$ for any $A \subset \mathcal{Z}$.

The data controller's objective is to obtain a representation mapping t that minimizes the statistical dependence between representation Z and sensitive attribute S. Therefore, for any test $f : \mathcal{Z} \to \{0, 1\}$ that decides whether the class conditional distributions $\mu_t^0 = P(Z|S = 0)$ and $\mu_t^1 = P(Z|S = 1)$ are identical, the data controller would like to minimize the discrepancy

$$\Delta(f, t) \triangleq |E_{x \sim \mu_t^1}[f(x)] - E_{x \sim \mu_t^0}[f(x)]|, \qquad (1)$$

where we make the dependence of Δ on representation mapping t explicit. In the context of fair machine learning, the test function f is either an auditor used by the data controller to estimate the statistical dependence between Z and S (function a in Figure 1); or, a classifier used by a data processor (function h in Figure 1) and $\Delta(f, t)$ then measures the demographic parity of f (see Hardt et al. (2016)):

Definition 2.1. Demographic parity *Consider a representation distribution μ_t induced by a representation mapping $t : \mathcal{X} \to \mathcal{Z}$. A classifier $f : \mathcal{Z} \to \{0, 1\}$ used by a data processor satisfies $\delta-$ Demographic Parity on μ_t if and only if $\Delta(f, t) \leq \delta$.*

Since the data controller does not know ex-ante which classifier data processors will use, it has to construct a mapping t such that all classifiers $f : \mathcal{Z} \to \{0, 1\}$ satisfy $\delta-$ demographic parity on μ_t for some prespecified $\delta > 0$. A demographic parity certificate is therefore an upper bound on the demographic disparity of any classifiers that access samples from the representation distribution μ_t.

Definition 2.2. Demographic Parity Certificate *Let $\delta \geq 0$. A representation space $(\mathcal{Z}, \mu_t)$ can be certified with $\delta-$ demographic parity if and only if*

$$\Delta^*(t) \triangleq \sup_{f : \mathcal{Z} \to \{0,1\}} \Delta(f, t) = \delta. \qquad (2)$$

To construct a representation mapping certified with $\Delta^*(t)-$ demographic parity, the data controller needs to evaluate the supremum over all test functions/auditors f_n that are constructed on the basis of a finite sample $\mathcal{D}_n = \{(x_i, s_i)\}_{i=1}^n$. Let $\mathcal{F}_n$ denote the set of all auditors $f_n : \mathcal{Z} \times (\mathcal{Z} \times \{0, 1\})^n \to \{0, 1\}$ constructed from a sample of size n.

Definition 2.3. Empirical Demographic Parity Certificate *Let $n \geq 1$ and $\delta \geq 0$. A representation space $(\mathcal{Z}, \mu_t)$ is certified with an empirical $\delta-$ demographic parity certificate if and only if*

$$\Delta_n(t) \triangleq \sup_{f_n : \in \mathcal{F}_n} \Delta(f_n, t) = \delta. \qquad (3)$$

The question is how to choose a representation mapping $t : \mathcal{X} \to Z$ so that empirical certificates $\Delta_n(t)$ approximate well the true demographic parity certificate $\Delta^*(t)$. Approximation properties of empirical certificates are important for a data controller to upper bound the demographic disparity of any downstream processor that uses fresh samples obtained after t has been constructed.

Since the data controller cannot constrain the data distribution over $\mathcal{X} \times \{0, 1\}$, we are looking for distribution-free approximation rates. In general, distribution-free rates do not exist (Devroye et al. (2013), ch. 7). But, in our setting, the data controller has some control over the representation distribution via t. In fact, the approximation $\Delta^*(t) - \Delta_n(t)$ depends on how much information in X is encoded by t in Z. If t randomly maps $\mathcal{X}$ to $\mathcal{Z}$, the data controller can certify μ_t with $0-$ demographic parity, but μ_t is useless to downstream data processors. The data controller trades-off representation demographic parity with information by learning an encoder $t : \mathcal{X} \to \mathcal{Z}$ and a decoder function $g : \mathcal{Z} \to \mathcal{X}$ that solves the following fair empirical representation problem

$$\min_{t,g} \mathcal{L}_{rec}(g, t, \mathcal{D}_n) \text{ subject to } \Delta_n(t) \leq \delta, \qquad (4)$$

where $\delta > 0$ is a pre-specified demographic parity threshold and $\mathcal{L}_{rec}$ is a reconstruction loss.

2.2 Necessary Condition

This section identifies a necessary condition on t for an empirical demographic parity certificate to approximate $\Delta^*(t)$ well. The necessary condition bounds the amount of information measured by the χ^2 mutual information between feature X and representation Z:

$$I_{\chi^2}(X, Z) \triangleq E_x E_z \left(\frac{\mu_t(z) - \mu_t(Z|X = x)}{\mu_t(z)} \right)^2. \qquad (5)$$

The χ^2 mutual information relies on a statistical distance, the χ^2-divergence – $\chi^2(Z, Z|X) = \int_z \left(dP(Z|X)/dP(Z) - 1\right)^2 dP(Z)$ – to average the distance between Z and $Z|X = x$ for $x \in \mathcal{X}$. It has been used in information theory to estimate the information that flows through a neural network (see Goldfeld et al. (2019)). In the context of fair representation learning, we find that empirical demographic parity certificates cannot provide good approximations of the representation's true demographic parity if the χ^2 input-output mutual information is large:

Theorem 2.1. *Let $n \geq 1$. Consider a representation function $t : \mathcal{X} \rightarrow \mathcal{Z}$. Then, for all test function $f_n \in \mathcal{F}_n$*

$$\sup_{\mu} E_{\mathcal{D}_n} |\Delta^*(t) - \Delta(f_n, t)| \geq \sup_{\mu_x} \left(1 - \frac{1}{I_{\chi^2}(X, Z)}\right)^n, \tag{6}$$

where the suppremum on the left hand side is taken over all distributions μ over $\mathcal{X} \times \mathcal{S}$ and the suppremum on the right hand side is taken over all distribution μ_x over $\mathcal{X}$.

Encoding more information of X in Z exposes the representation distribution μ_t to mirroring distributions over $\mathcal{X}$ with heavy tails. Intuitively, μ_t is a (possibly infinite) mixture of conditional distributions $P(Z|X = x)$ for $x \in \mathcal{X}$ and $I_{\chi^2}(X, Z)$ measures an average distance between those conditional distributions. As $I_{\chi^2}(X, Z)$ increases, the conditional distributions $P(Z|X = x)$ become far apart for a growing mass of $x \in \mathcal{X}$. It generates a representation distribution too complex for a finite sample to represent it and for an auditor f_n to detect all the correlations between representation and sensitive attribute.

Theorem 2.1 implies a trade-off between the information passed from features to representations and the approximation rate of empirical demographic parity certificates:

Corollary 2.1. *With the notations from Theorem 2.1,*

- *If* $\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n} |\Delta^*(t) - \Delta(f_n, t)| \leq \epsilon_n$, *then for all distributions over the feature space $\mathcal{X}$, $I_{\chi^2}(X, Z) \leq \frac{1}{1 - \epsilon_n^{\frac{1}{n}}}$.*
- *If there exists a distribution over $\mathcal{X}$ such that $I_{\chi^2}(X, Z) = \infty$,*

$$\inf_{f_n \in \mathcal{F}_n} \sup_{\mu} \Delta^*(t) - \Delta(f_n, t) \geq 1. \tag{7}$$

For the approximation rate of $\Delta^*(t) - \Delta(f_n, t)$ to be $O(n^{-s})$ for some $s > 0$, it is necessary for the χ^2 mutual information between feature and representation to be bounded above by $O(n/(s \ln(n))$ for *all* distributions over $\mathcal{X}$. On the other hand, representation functions t for which the χ^2 mutual information is infinite for some distribution over the features space, never guarantee a meaningful approximate rate between $\Delta^*(t)$ and $\Delta_n(f_n, t)$.

Examples: The results in corollary 2.1 imply that empirical certificates of representation distributions induced by many common encoders do not have meaningful approximation rates:

- Suppose that t is injective from $\mathbb{R}^D$ to $\mathbb{R}^d$. Then, there exists a distribution over $\mathcal{X} \times \{0, 1\}$ such that $I_{\chi^2}(X, Z) = \infty$ and thus, $\Delta^*(t) = 1$, but $\Delta(f_n, t) = 0$ for all auditing functions f_n.
- Suppose that $|\{t(x)|x \in \mathcal{X}\}| \geq n/(\ln(n))^{\alpha}$, for some $\alpha < 1$. Then, the approximation rate of $\Delta(f_n, t)$ for all auditing functions f_n is $\omega(n^{-s})$ for any $s > 0$.

2.3 Sufficient Condition

This section shows that a finite χ^2 mutual information between feature and representation for all distributions over $\mathcal{X}$ is a sufficient condition for empirical demographic parity certificates to converge at a $O(n^{-1/2})$ rate.

Theorem 2.2. *Let $n \geq 1$. Consider a representation mapping $t : \mathcal{X} \rightarrow \mathcal{Z}$. Then, for all distribution μ over $\mathcal{X} \times \{0, 1\}$ with where $n_s = |\{i | s_i = s\}|$ and for all $f_n \in \mathcal{F}_n$*

$$E_{\mathcal{D}_n} |\Delta^*(t) - \Delta(f_n, t)| \leq 2 \sum_{s=0,1} \sqrt{\frac{I_{\chi^2}(X, Z|S = s)}{n_s}}$$

A finite χ^2 mutual information between X and Z implies that $p(Z)$ and $p(Z|X)$ are close in the sense of the χ^2 divergence and thus by sampling representations from $P(Z|X)$, we have a non-zero probability to sample all the atoms that can form the representation distribution μ_t and thus to detect all the dependence between representations and sensitive attributes.

2.4 χ^2 - versus Classic Mutual Information

Our results in Theorems 2.1 and 2.2 highlight a connection between χ^2 mutual information and approximation rate of empirical certificates. A similar result cannot be obtained with the classic mutual information $I_{Sh}(X, Z)$ that is based on Shannon entropy.

To demonstrate this point, we construct the following distribution μ over $\mathcal{X} \times \{0, 1\}$. Features are uniformly distributed over $[0, 1]$ and $t(x) = i$ for $x \in [1/i, 1/(i + 1))$ and $i > 0$. For each $i > 0$, the sensitive attribute is constant over $[1/i, 1/(i + 1))$ and equal to 1 with probability 1/2. We show in the appendix that $I_{Sh}(X, Z) < \ln(2)/2 + 2$, but $I_{\chi^2}(X, Z) = \infty$. Since the sensitive attribute S is a deterministic function of the representation $Z = t(X)$, $\Delta^*(t) = 1$. But, for a finite sample of size n, $E_{\mathcal{D}_n} \Delta(f_n, t)$ is zero for all auditors f_n, despite $I_{Sh}(X, Z) < \infty$.

3 Smooth and Fair Representations

The previous section suggests restricting the fair representation problem (4) to encoder t for which the

χ^2-mutual information between feature and representation is finite for all distributions over $\mathcal{X}$. Here, we meet this condition by adding an additive Gaussian white noise (AGWN) channel to the encoder. For any representation mapping $t : \mathcal{X} \rightarrow \mathcal{Z}$, we denote t_σ the convolution of t with a Gaussian noise $\mathcal{N}(0, \sigma^2 I_d)$: $t_\sigma(X) = t(X) + noise$, with $noise \sim \mathcal{N}(0, \sigma^2 I_d)$.

3.1 Convergence of Smoothed Empirical Certificate

The convolved representation $Z_\sigma = Z + noise$ generated by t_σ has a distribution denoted $\mu_{t*\sigma}$. The convolution smoothes the representation distribution by making $P(Z_\sigma|X)$ a Gaussian whose support covers the support of the representation distribution $P(Z_\sigma)$ and thus, guarantees that samples from different conditional distributions $P(Z_\sigma|X = x)$ are not too far away.

Theorem 3.1. *Let $\sigma > 0$ and $n \geq 1$. For all representation mapping $t : \mathcal{X} \rightarrow \mathcal{Z}$ and for any distribution over $\mathcal{X}$, if $||t||_\infty \triangleq \sup_{x\in\mathcal{X}} ||t(x)||_2$, then for $s \in \{0, 1\}$*

$$I_{\chi^2}(X, Z|S = s) \leq \exp\left(\frac{||t||_\infty^2}{\sigma^2}\right) < \infty. \tag{8}$$

Therefore,

$$\begin{aligned}&\inf_{f_n\in\mathcal{F}_n} \sup_{\mu} E_{\mathcal{D}_n}[\Delta^*(t_\sigma) - \Delta(t_\sigma, f_n)] \\ &\leq 2\exp\left(\frac{||t||_\infty^2}{2\sigma^2}\right)(n_0^{-1/2} + n_1^{-1/2}).\end{aligned} \tag{9}$$

The upper bound in Theorem 3.1 does not depend on the dimensions d of the representation space $\mathcal{Z}$, but only on $n^{-1/2}$ and on the ratio $||t||_\infty/\sigma$ that can be interpreted as a signal-to-noise ratio in the AGWN channel. Larger values of $||t||_\infty$ increase the variance of Z and thus require larger noise σ to keep the conditional distribution $P(Z_\sigma|X)$ close to the distribution $P(Z_\sigma)$. The bound is only meaningful if $||t||_\infty < \infty$, which holds, for example, if the features space is bounded and t is a continuous mapping.

Both Theorems 2.2 and 3.1 rely on a plug-in auditor that first estimates the class-conditional densities $\mu_{t*\sigma}^0$ and $\mu_{t*\sigma}^1$. From a sample $\mathcal{D}_n = \{(x_i, s_i)\}_{i=1}^n$, we construct an empirical estimate of $\mu_{t*\sigma}$ over $\mathcal{Z} \times \{0, 1\}$ as

$$\mu_{n,\sigma}(z, s) = \frac{1}{n}\sum_{i=1, s_i=s}^{n} P(z|X = x_i) \tag{10}$$

with $P(.|X = x_i) \sim \mathcal{N}(t_n(x_i), \sigma I_d)$. Our plug-in auditor f_n^{plug} compares $\mu_{n,\sigma}(z, 0)$ to $\mu_{n,\sigma}(z, 1)$:

$$f_n^{plug}(z) = \begin{cases} 0 & \text{if } \mu_{n,\sigma}(z, 0) \geq \mu_{n,\sigma}(z, 1) \\ 1 & \text{otherwise.}\end{cases} \tag{11}$$

Since we obtain the upper bounds in Theorems 2.2 and 3.1 with the plug-in auditor f_n^{plug}, we can guarantee that the representation demographic parity is within $O(n^{-1/2})$ of the empirical certificate signed by f_n^{plug}.

3.2 Learning Fair Representation

In practice, the representation mapping t and the decoder g are modelled by neural networks. An AGWN channel is added to t to learn a smoothed representation distribution $\mu_{t*\sigma}$. The data controller trades off minimizing a reconstruction loss $\mathcal{L}_{rec}(t, g) = E_x[l_{rec}(t, g, x)]$ with minimizing demographic disparity $\mathcal{L}_{DP}(t) = \Delta^*(t_\sigma)$. With a sample $\mathcal{D}_n = \{(x_i, s_i)\}_{i=1}^n$, the data controller uses the plug-in auditor and solves the empirical minimization problem as

$$\min_{t,g} \frac{1}{n}\sum l_{rec}(t, g, x_i) + \lambda\Delta(f_n^{plug}, t_\sigma), \tag{12}$$

where λ controls for the strength of the fairness constraint imposed on the representation distribution. The minimization problem in (12) differs from previous work on fair representation learning because of the noise added to Z and thus, provides theoretical guarantees that $\Delta(f_n^{plug}, t_\sigma,)$ approximates $\Delta^*(t_\sigma)$ at a rate $O(n^{-1/2})$.

Moreover, the empirical demographic parity certificate can be computed without modelling the auditor by an additional adversarial neural network. This is because we can use our empirical estimates (10) of the class-conditional densities to estimate the posterior distribution $\eta(z, s) = P(S = s|Z = z)$ as $\eta_n(z, s) = \mu_{n,\sigma}(z|S = s)/\mu_{n,\sigma}(z)$, where $\mu_{n,\sigma}(z) = \mu_{n,\sigma}(z, 1) + \mu_{n,\sigma}(z, 0)$. Since $\Delta^*(t)$ relates to the balanced error rate of predicting the sensitive attributes (see proof of 2.2 or Feldman et al. (2015)), we can write $\Delta^*(t) = \mathcal{L}_{DP}(\mu_{t,\sigma})$, where $\mathcal{L}_{DP}(\mu_{t,\sigma}) = E_{z\sim\mu_{t,\sigma}}[|\eta(z, 1) - \eta(z, 0)|]$ (see Zhao et al. (2013)). Our approach relies on two results: (i) for any finite sample of size n, $\mathcal{L}_{DP}(\mu_{n,\sigma})$ approximates well $\mathcal{L}_{DP}(\mu_{t*\sigma})$; (ii) $\mathcal{L}_{DP}(\mu_{n,\sigma})$ can be estimated efficiently by Monte-Carlo estimation. The first observation uses the following result, which is a consequence of Theorem 3.1

Theorem 3.2. *Let $\sigma > 0$ and $n \geq 1$. For all representation mapping $t : \mathcal{X} \rightarrow \mathcal{Z}$*

$$\begin{aligned}&\sup_{\mu} E_{\mathcal{D}_n}|\mathcal{L}_{DP}(\mu_{t*\sigma}) - \mathcal{L}_{DP}(\mu_{n,\sigma})| \\ &\leq 2\exp\left(\frac{||t||_\infty^2}{2\sigma^2}\right)(n_0^{-1/2} + n_1^{-1/2}).\end{aligned} \tag{13}$$

Therefore, we can use $\mathcal{L}_{DP}(\mu_{n,\sigma})$ as an approximation of $\mathcal{L}_{DP}(\mu_{t*\sigma})$. That is, in place of $\mu_{t,\sigma}$, we propose to use the distribution $\mu_{n,\sigma}$, for which η_n is the posteriori probability. Moreover, $\mathcal{L}_{DP}(\mu_{n,\sigma})$ can be efficiently

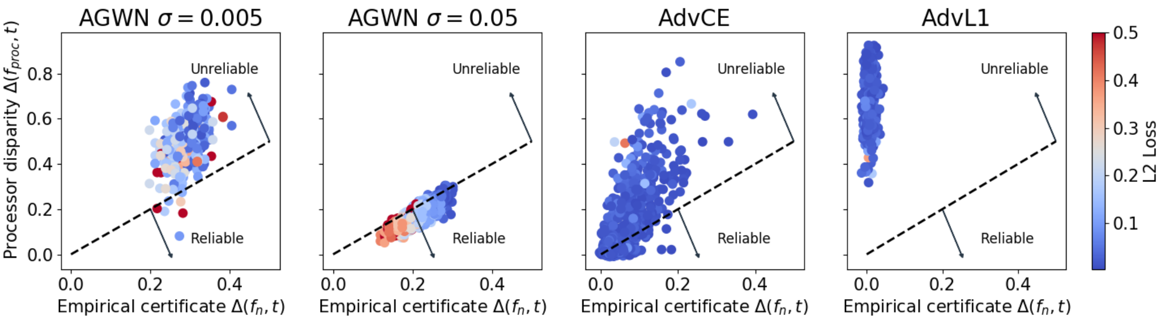

Figure 2: Generalization of empirical demographic parity certificates for the Swiss Roll data. Dots are colored by reconstruction loss.

approximated by Monte Carlo integration. For a sample $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, $\mu_{n,\sigma}^0$ and $\mu_{n,\sigma}^1$ are mixtures of d-dimensional Gaussians. Thereby, we approximate $\mathcal{L}_{DP}(\mu_{n,\sigma})$ with

$$\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} E_\epsilon[|\eta_n(z_{ij}, 1) - \eta_n(z_{ij}, 0)|, \tag{14}$$

where $z_{ij} = t(x_i) + noise_{ij}$, $\{noise_{ji}\}$ is a vector of $n \times m$ draws from a d-dimensional Gaussian $\mathcal{N}(0, \sigma I_d)$ and m is the number of draws per sample point. $\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma})$ is an unbiased approximation of $\mathcal{L}_{DP}(\mu_{n,\sigma})$ and achieves a Mean-Squared-Error (MSE) of order $O(n^{-1}m^{-1})$ (see proof of Theorem 4 in appendix).

To sum up, the data controller learns (t, g) by minimizing the following combined empirical loss

$$\min_{\theta,\varphi} \frac{1}{n}\sum_i l_{rec}(t, g, x_i) + \lambda\hat{\mathcal{L}}_{DP}(\mu_{n,\sigma}). \tag{15}$$

Practical implementation. We minimize the loss (15) by stochastic gradient descent. Each mini-batch is split in half: the first half is used to estimate $\mu_{n,\sigma}$ as in (10); the second half to estimate the loss in (15). At the end of training, we compute a leave-one-out balanced error rate $BER(f_n^{plug})$ for the plug-in auditor on both a test and train samples and infer an empirical certificate as $\Delta(f_n^{plug}, t) = 1 - 2BER(f_n^{plug})$ (see Feldman et al. (2015)).

Choice of σ. The Gaussian noise σ is an hyper-parameter chosen so that empirical certificates estimated on train and test data are similar. To set the value σ, we divided the data in training/validation/test sets. The test set is used for evaluating the disparity of downstream classifiers. The validation set is used to tune the value of σ as follows: we start with a small value of σ ($\sigma \approx 0.005$) and increase it until the value of empirical certificate $\Delta(f_n^{plug}, t)$ estimated by our plug-in auditor f_n^{plug} on the training set exceeds (up to a small tolerance factor 0.025) the one estimated on the validation set. Therefore, we choose σ so that the plug-in auditor generalizes well to unseen data.

4 Experiments

The objective of this experimental section is to demonstrate that (i) our AGWN fair representation method, unlike competitive approaches, generates robust fairness certificates that generalize to unseen data; and, (ii) it is competitive with existing fair representation methods in terms of fairness-accuracy trade-off. All details related to dataset and neural network architectures are in the appendix.

Datasets. We first consider two synthetic datasets. Our first synthetic data consists of two $3D$ Swiss rolls: one for $S = 0$ and one shifted South-West for $S = 1$. Our second synthetic data is a variant of the DSprites dataset (Matthey et al. (2017)) that contains 64 by 64 black and white images of various shapes (heart, square, circle). The DSprites dataset has six independent factors of variation: color (black or white); shape (square, heart, ellipse), scales (6 values), orientation (40 angles in $[0, 2\pi]$); x- and y- positions (32 values each). We adapt the sampling to generate a source of potential unfairness as in Creager et al. (2019).

We also apply our approach of fair representation learning with a AGWN channel to two fair learning benchmarks, Adults[1] and Heritage[2]. The Adults dataset contains $49K$ individuals and includes information on 10 features related to professional occupation, education attainment, race, capital gains, hours worked and marital status. The sensitive attribute is the gender to which individuals self-identify to.

The Health Heritage dataset contains $220K$ individuals with 66 features related to age, clinical diagnoses and procedure, lab results, drug prescriptions

[1] https://archive.ics.uci.edu/ml/datasets/adult
[2] https://foreverdata.org/1015/index.html

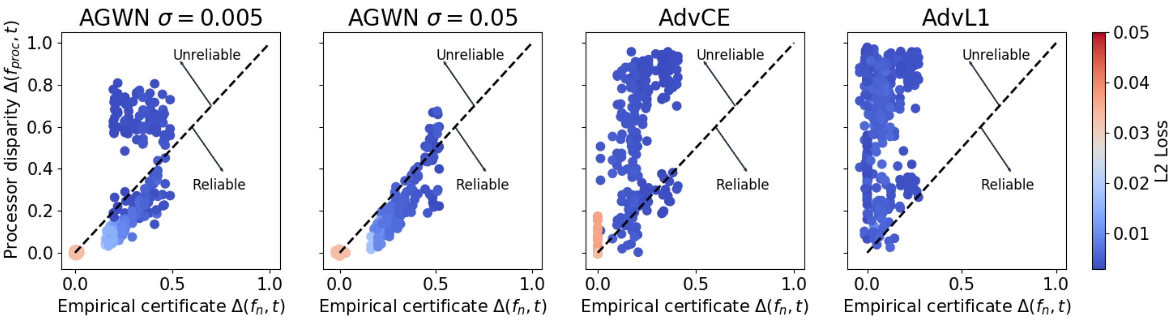

Figure 3: Generalization properties of empirical demographic parity certificates for DSprites. See Figure 2.

and claims payment aggregated over 3 years. The sensitive attribute is the gender to which individuals self-identify to.

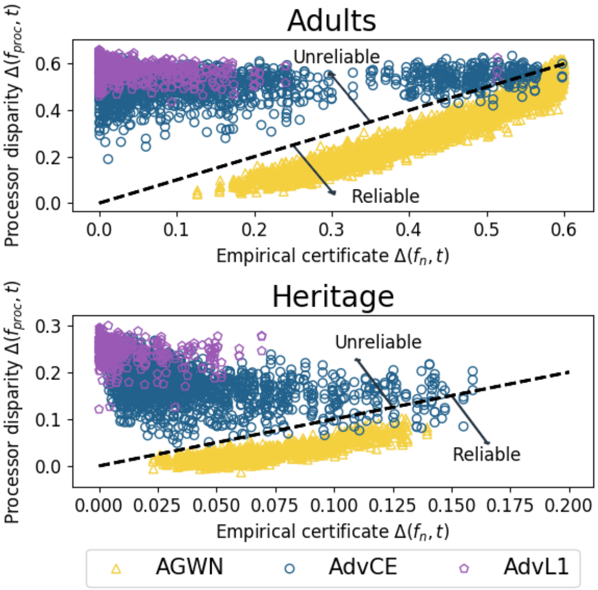

Figure 4: Generalization of empirical demographic parity certificates for Adults and Heritage.

Effect of noise on certificate reliability. We train an encoder-decoder mapping (t, g) with an increasing amount of Gaussian noise; estimate an empirical $\Delta(f_n^{plug}, t)-$demographic parity certificate; and then, test whether $\Delta(f_n^{plug}, t)$ is larger than the demographic disparity $\Delta(f_{proc}, t)$ of different downstream processors f_{proc} that predict sensitive attributes from new samples of the representation distribution. Empirical certificates are robust if $\Delta(f_n^{plug}, t) \geq \Delta(f_{proc}, t)$ for any of the processors f_{proc}.

All datasets are split into a train set for training the auto-encoder (t, g); two test sets to first train downstream processors and then evaluate their accuracy.

Comparative adversarial methods. We benchmark the use of an AGWN channel with approaches in fair representation learning based on adversarial auditor trained with (i) a cross-entropy loss (AdvCE, Edwards and Storkey (2015)); or, with (ii) a group L1 loss (AdvL1, Madras et al. (2018)). The processors have same width and depth as the adversarial auditors.

5 Results and Discussion

Certificate reliability. Figure 2 and Figure 3 provide a comparison of AdvCE, AdvL1 and AGWN with respect to the generalization of the empirical demographic parity certificates for the Swiss roll and Dspirites datasets, respectively. Each dot shows empirical demographic parity certificate $\Delta(f_n, t)$ for an encoder $t \in \{AGWN, AdvCE, AdvL1\}$ against an estimate of the disparity $\Delta(f_{proc}, t)$ of downstream processors predicting sensitive attributes.

Figure 2 and Figure 3 show that the AGWN channel improves how empirical certificates approximate the demographic parity of the representation distribution. As the Gaussian noise σ increases from $\sigma = 0.005$ to $\sigma = 0.05$, the $\Delta(f_n^{plug}, t)$ empirical certificate upper bounds the demographic disparity $\Delta(f_{proc}, t)$ for any of the downstream processors we built, regardless of their complexity. Moreover, the variance of $\Delta(f_n, t) - \Delta(f_{proc}, t)$ decreases as the Gaussian noise increases. This is consistent with the upper bound in Theorem 3.1, which decreases with smaller signal-to-noise ratio $||t||_\infty/\sigma$.

Comparative adversarial approaches. Figure 2 and Figure 3 also show that for both comparative methods, the empirical certificate $\Delta(f_{adv}, t)$ estimated by the adversarial auditor underestimates significantly the disparity obtained by downstream processors on fresh samples from the representation distribution. For example, for the Swiss Roll dataset, 18.2% of near zero empirical certificates $(\Delta(f_{adv}, t) \leq 0.1)$ do not preclude a processor's disparity larger than 0.3.

Real world data. In Figure 4, we compare again

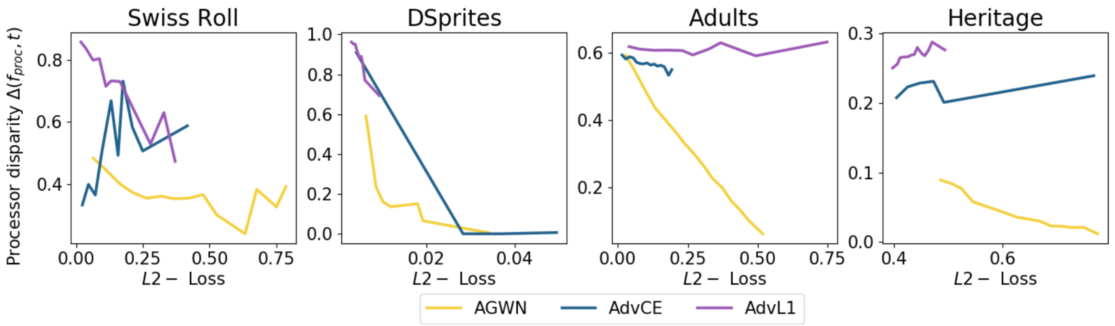

Figure 5: Reconstruction loss v.s. worst disparity attained by downstream processors.

disparity of downstream processors against empirical certificate on Adults and Heritage. Certificate's reliability is measured by the position of the scatter plot relative to the 45° line. Figure 4 confirms that (i) the AGWN channel is sufficient for empirical certificates to upper-bound the demographic disparity obtained by various downstream processors; and, (ii) that comparative methods (AdvCE, AdvL1) generate empirical fairness certificates that do not bound the disparity of downstream processors.

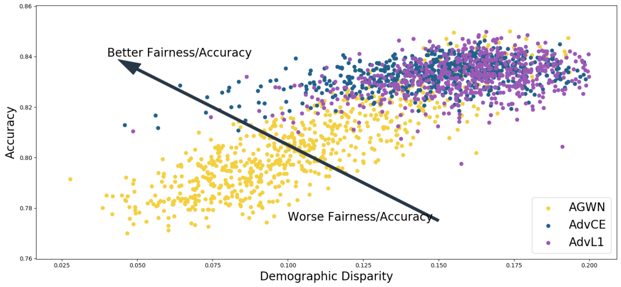

Figure 6: Accuracy-fairness trade-off.

Accuracy-fairness trade-off. For the Swiss Roll dataset (Figure 2), AGWN's reliability appears to come at the cost of a larger reconstruction loss for a given empirical fairness certificate. However, it is not the case for DSprites. Moreover, a fair comparison across methods requires to measure reconstruction loss against the worst disparity attained by a downstream processor, i.e. the upper bound of the point clouds in Figure 2 and 3. Figure 5 plots the 95^{th} quantile of the demographic disparity of downstream processors for a given reconstruction loss. It shows that across all datasets, for a given $L2-$loss, the worst demographic disparity of downstream processors is lower when the representations are generated by AGWN than AdvCE or AdvL1. Moreover, for Swiss Roll and Adults, larger reconstruction losses (≥ 0.5 for Swiss Roll; ≥ 0.25 for Adults) with AGWN correspond to low levels of processors' disparity that are never reached by comparative methods.

To explore further how the AGWN channel affects the information contained in the representation, we compare the demographic disparity and the accuracy of downstream processors that predict a task label Y. We retrain the three fair learning methods – AdvCE, AdvL1 and AGWN – on the Adults dataset but leave out the income feature. We map test samples into their corresponding representations and predict whether their income is over $50K$. In Figure 6, we sweep the parameter space for different values of the fairness constraint λ in (12). Each dot compares the accuracy and the demographic disparity of neural networks of various depth and width. The higher the accuracy of downstream processors for a given level of disparity, the better the fairness-accuracy tradeoff. We can draw two conclusions from this experiment. First, for level of disparity between 0.10 and 0.20, AGWN offers the same fairness-accuracy tradeoff as AdvL1 or AdvCE. Second, our AGWN method is the only one for which varying the coefficient on the fairness constraint allows to systematically reach low level of disparity (≤ 0.1). Consistent with Figure 5, very few simulations of AdvCe and AdvL1 lead to the demographic disparity of downstream processors to be less than 0.075, regardless of the strength of the fairness penalty used during the training of the autoencoder. Although the AGWN channel limits the maximum amount of information that is transferred from the data to the representation (see Cover and Thomas (2012)), it also allows for a better empirical approximation of demographic parity and thus helps guiding the representation mapping toward the correct fairness-information trade-off.

6 Conclusion

This paper investigates whether a data controller could generate representations of the data with fairness guarantees that would hold for any downstream processor using samples from the representation distribution. For demographic parity certificate to approximate well the demographic parity of all future data processors it is necessary and sufficient to bound the χ^2 mutual information between feature and representation. We meet this condition by adding an AGWN channel while learning a fair representation of the data.

Our work opens promising research avenues in fair representation learning. An AGWN channel is only one of many approaches to bound the χ^2 mutual information between feature and representation and ensure the reliability of the learned representations.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. (2018). A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Friedler, S. A. and Wilson, C., editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Calders, T. and Žliobaitė, I. (2013). Why unbiased computational processes can lead to discriminative decision procedures. In *Discrimination and privacy in the information society*, pages 43–57. Springer.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc.

Chouldechova, A. and Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. S. (2019). Flexibly fair representation learning by disentanglement. *ArXiv*, abs/1906.02589.

Devroye, L., Györfi, L., and Lugosi, G. (2013). *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407.

Edwards, H. and Storkey, A. (2015). Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM.

Gardner, J., Brooks, C., and Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th International Conference on Learning Analytics & Knowledge*, pages 225–234. ACM.

Gitiaux, X. and Rangwala, H. (2019). mdfa: Multi-differential fairness auditor for black box classifiers. In *IJCAI*.

Goldfeld, Z., Greenewald, K., Polyanskiy, Y., and Weed, J. (2019). Convergence of smoothed empirical measures with applications to entropy estimation. *arXiv preprint arXiv:1905.13576*.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2018). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning*, pages 2569–2577.

Kim, M., Reingold, O., and Rothblum, G. (2018). Fairness through computationally-bounded awareness. In *Advances in Neural Information Processing Systems*, pages 4842–4852.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The variational fair autoencoder.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Oneto, L., Donini, M., Maurer, A., and Pontil, M. (2019). Learning fair and transferable representations.

Pfohl, S., Marafino, B., Coulet, A., Rodriguez, F., Palaniappan, L., and Shah, N. H. (2019). Creating fair models of atherosclerotic cardiovascular disease risk. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 271–278. ACM.

ProPublica (2016). How we analyzed the compas recidivism algorithm. *ProPublica*.

Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, pages 325–333.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Zhao, M.-J., Edakunni, N., Pocock, A., and Brown, G. (2013). Beyond fano's inequality: bounds on the optimal f-score, ber, and cost-sensitive risk and their implications. *Journal of Machine Learning Research*, 14(Apr):1033–1090.