Fair Representations by Compression

Xavier Gitiaux, Huzefa Rangwala
George Mason University
xgitiaux@gmu.edu, rangwala@gmu.edu

Abstract

Organizations that collect and sell data face increasing scrutiny for the discriminatory use of data. We propose a novel unsupervised approach to transform data into a compressed binary representation independent of sensitive attributes. We show that in an information bottleneck framework, a parsimonious representation should filter out information related to sensitive attributes if they are provided directly to the decoder. Empirical results show that the proposed method, FBC, achieves state-of-the-art accuracy-fairness trade-off. Explicit control of the entropy of the representation bit stream allows the user to move smoothly and simultaneously along both rate-distortion and rate-fairness curves.

1 Introduction

A growing body of evidence has questioned the fairness of machine learning algorithms across a wide range of applications, including judicial decisions (ProPublica 2016), face recognition (Buolamwini and Gebru 2018), degree completion (Gardner, Brooks, and Baker 2019) or medical treatment (Pfohl et al. 2019). Of particular concerns are potential discriminatory uses of data on the basis of racial or ethnic origin, political opinion, religion, or gender.

Therefore, organizations that collect and sell data are increasingly liable if future downstream uses of the data are biased against protected demographic groups. One of their challenges is to anticipate and control how the data will be processed by downstream users. Unsupervised fair representation learning approaches (Madras et al. (2018), Zemel et al. (2013), Gitiaux and Rangwala (2020), Moyer et al. (2018)) offers a flexible fairness solution to this challenge. A typical architecture in fair representation learning includes an encoder that maps the data into a representation and a decoder that reconstructs the data from its representation. The objective of the architecture is to extract from a data $X$ the underlying latent factors $Z$ that correlate with unobserved and potentially diverse task labels, while remaining independent of sensitive factors $S$.

This paper asks whether an encoder that filters out information redundancies could generate fair representations. Intuitively, if sensitive attributes $S$ are direct inputs to the decoder, an encoder that aims for conciseness would not waste code length to encode information related to $S$ in the latent factors $Z$. We show that in an information bottleneck framework (Tishby, Pereira, and Bialek 2000), this intuition is theoretically founded: constraining the information flowing from the data $X$ to the representation $Z$ forces the encoder to control the dependencies between sensitive attributes $S$ and representations $Z$.

It is sufficient to constraint the mutual information $I(Z, X)$ between $Z$ and $X$ in order to minimize the mutual information $I(Z, S)$ between $Z$ and $S$.

Therefore, instead of directly penalizing $I(Z, S)$, we recast fair representation learning as a rate distortion problem that controls explicitly the bit rate $I(Z, X)$ encoded in the latent factors $Z$. We model the representation $Z$ as a binary bit stream, which allows us to monitor the bit rate more effectively than floating point representations that may maintain redundant bit patterns. We estimate the entropy of the code $Z$ with an auxiliary auto-regressive network that predicts each bit in the latent code $Z$ conditional on previous bits in the code. One advantage of the method is that the auxiliary network collaborates with the encoder to minimize the cross-entropy of the code.

Empirically, we demonstrate that the resulting method, Fairness by Binary Compression (henceforth, FBC) is competitive with state-of-the art methods in fair representation learning. Our contributions are as follows:

1. We show that controlling for the mutual information $I(Z, X)$ is an effective way to remove dependencies between sensitive attributes and latent factors $Z$, while preserving in $Z$, the information useful for downstream tasks.

2. We find that compressing the data into a binary code as in FBC generates a better accuracy-fairness trade-off than limiting the information channel capacity by adding noise (as in variants of $\beta$-VAE, Higgins et al. (2017)).

3. We show that increasing the value of the coefficient on the bit rate constraint $I(Z, X)$ in our information bottleneck framework allows to move smoothly along both rate-distortion and rate-fairness curves.

Related work. The machine learning literature increasingly explores how algorithms can adversely impact protected demographic groups (e.g individuals self-identified as Female or African-American) (see Chouldechova and Roth (2018) for a review). Research questions revolve around how...
to define fairness (Dwork et al. [2012]), how to enforce fairness in standard classification algorithms (e.g. Agarwal et al. [2018], Kim, Reingold, and Rothblum [2018], Kearns et al. [2018]) or audit a black box classifier for its fairness (e.g. Feldman et al. [2015], Gitaiaux and Rangwala [2019]).

This paper relates to recent efforts towards transforming data into fair and general purpose representations that are not tailored to a pre-specified specific downstream task. Many contributions use a supervised setting where the downstream task label is known while training the encoder-decoder architecture (e.g. Madras et al. [2018], Edwards and Storkey [2015], Moyer et al. [2018], Song et al. [2018] or Jaiswal et al. [2019]). However, Zemel et al. [2013], Gitaiaux and Rangwala [2020] and Locatello et al. [2019] argue that in practice, an organization that collects data cannot anticipate what the downstream use of the data will be. In this unsupervised setting, the literature has focused on penalizing approximations of the mutual information between representations and sensitive attributes; maximum mean discrepancy penalty (Gretton et al. [2012]) for deterministic (Li, Swersky, and Zemel [2014]) or variational (Louizos et al. [2015]) autoencoders (see Table 1); cross-entropy of an adversarial auditor that predicts sensitive attributes from the representations (Madras et al. [2018], Edwards and Storkey [2015], Zhang, Lemoine, and Mitchell [2018] or Xu et al. [2018]).

Our approach contrasts with existing work since it does not control directly for the leakage between sensitive attributes and representations. FBC obtains fair representations only by controlling its bit rate. In a supervised setting, Jaiswal et al. [2019] show that nuisance factors can be removed from a representation by over-compressing it. We extend their insights to unsupervised settings and show the superiority of bit stream representations over noisy ones to remove nuisance factors. Our insights could offer an effective alternative to methods that learn representations invariant to nuisance factors (e.g. (Achille and Soatto 2018), (Jaiswal et al. 2019), (Jaiswal et al. 2018)).

Our paper borrows soft-quantization techniques when backpropagating through the model (Agustsson et al. 2017) and hard quantization techniques during the forward pass (Mentzer et al. 2018). We find that in our fair representation setting, explicit control of the bit rate of the representation leads to better accuracy-fairness trade-off than floating point counterpart. We estimate the entropy of the code as in (Mentzer et al. 2018) by computing the distribution $P(Z)$ of $Z$ as an auto-regressive product of conditional distributions, and by modeling the auto-regressive structure with a PixelCNN architecture (Oord, Kalchbrenner, and Kavukcuoglu 2016, Van den Oord et al. 2016).

## 2 Fair Information Bottleneck

Consider a population of individuals represented by features $X \in \mathcal{X} \subset [0, 1]^{d_x}$ and sensitive attributes in $S \in \mathcal{S} \subset [0, 1]^{d_s}$, where $d_x$ is the dimension of the feature space and $d_s$ is the dimension of the sensitive attributes space. In this paper, we do not restrict ourselves to binary sensitive attributes and we allow $d_s > 1$. The objective of fair representation learning is to map the features space $\mathcal{X}$ into a $m-$dimensional representation space $\mathcal{Z} \subset [0, 1]^m$, such that

- (i) $Z$ maximizes the information related to $X$, but (ii) minimizes the information related to sensitive attributes $S$. We can express this as

$$\max_Z I(X, \{Z, S\}) - \gamma I(Z, S)$$

where $I(X, S)$ and $I(X, \{Z, S\})$ denote the mutual information between $Z$ and $S$ and between $X$ and $(Z, S)$, respectively; and $\gamma \geq 0$ controls the fairness penalty $I(Z, S)$.

Existing methods focus on solving directly the problem (1) by approximating the mutual information $I(Z, S)$ between $Z$ and $S$ via the cross-entropy of an adversarial auditor that predicts sensitive attributes from the representations (Madras et al. [2018], Edwards and Storkey [2015], Gitaiaux and Rangwala [2020]) or via the maximum mean discrepancy between $Z$ and $S$ (Louizos et al. [2015]).

In this paper, we instead reduce the fair representation learning program (1) to an information bottleneck problem that consists of encoding $X$ into a parsimonious code $Z$, while ensuring that this code $Z$ along with a side channel $S$ allows a good reconstruction of $X$. The mutual information between $X$ and $S$ can be written as

$$I(Z, S) \overset{(a)}{=} I(Z, \{X, S\}) - I(Z, X|S)$$

$$\overset{(b)}{=} I(Z, X) + I(Z, S|X) - I(Z, X|S)$$

$$\overset{(c)}{=} I(Z, X) - I(Z, X|S)$$

$$\overset{(d)}{=} I(Z, X) - I(X, \{Z, S\}) + I(X, S).$$

where (a), (b) and (d) use the chain rule for mutual information; and, (c) uses the fact that $Z$ is only encoded from $X$, so $H(Z|X, S) = H(Z|X)$ and $I(Z, X|S) = H(Z|X) - H(Z|X,S) = 0$. Since the mutual information between $X$ and $S$ does not depend on the code $Z$, the fair representation learning (1) is equivalent to the following fair information bottleneck:

$$\max_Z (1 + \gamma) I(X, \{Z, S\}) - \gamma I(Z, X).$$

Intuitively, compressing information about $X$ forces the code $Z$ to avoid information redundancy, particularly redundancy related to the sensitive attribute $S$, since the decoder has direct access to $S$. Note that there is no explicit constraint in (2) to impose independence between $Z$ and $S$.

If the representation $Z$ is obtained by a deterministic function of the data $X$, once $X$ is known, $Z$ is known and $H(Z|X) = 0$. Therefore, the mutual information $I(Z, X)$ is equal to the entropy $H(Z)$ of the representation $Z$. Since the entropy of the data $X$ does not depend on the representation $Z$, we can replace $I(X, \{Z, S\})$ by $E_{x,z,s} \log(P(x|z, s))$ in the information bottleneck (2) and solve for:

$$\min_Z E_{x,z,s}[- \log(P(X|Z, S)) + \beta H(Z)],$$

where $\beta = \gamma/(\gamma + 1)$. Therefore, the fair representation problem, in its information bottleneck interpretation, can be recast as a rate-distortion trade-off. A lossy compression of the data into a representation $Z$ forces the independence
between sensitive attribute and representation but increases the distortion cost measured by the negative log-likelihood of the reconstructed data $E_{x,z}[-\log \mathbb{P}(x|z,S)]$. The parameter $\beta$ in equation (3) controls the competitive objectives of low distortion and fairness-by-compression: the larger $\beta$, the fewer the dependencies between $Z$ and $S$.

### 3 Proposed Method

There are two avenues to control for $I(Z,X)$ in the information bottleneck (2) (see Figure 1): (i) adding noise to $Z$ to control the capacity of the information channel between $X$ and $Z$; or, (ii) storing $Z$ as a bit stream whose entropy is explicitly controlled.

The noisy avenue (i) is a variant of variational autoencoders, so called $\beta$-VAE (Higgins et al. 2017), that models the posterior distribution $P(Z|X)$ of $Z$ as Gaussian distributions (see Figure 1a). The channel capacity and thus the mutual information between $X$ and $Z$ is constrained by minimizing the Kullback divergence between these posterior distributions and an isotropic Gaussian prior (Braithwaite and Klein 2018). In the context of fair representation learning, Louizos et al. (2015) and Creager et al. (2019) use variants of $\beta$-VAE, but do not focus on how limiting the channel capacity $I(Z,X)$ could lead to fair representations. Instead, they add further constraints on $I(Z,S)$.

We implement the binary avenue with a method $\gamma$-FBC (see Figure 1b) that consists of an encoder $F: \mathcal{X} \rightarrow \mathbb{R}^m$, a binarizer $B: \mathbb{R}^m \rightarrow \{0,1\}^m$ and a decoder $G: \{0,1\}^m \times S \rightarrow \mathcal{X}$. The encoder $F$ maps each data point $x$ into a latent variable $e = F(x)$, the binarizer $B$ binarizes the latent variable $e$ into a bit stream $z$ of length $m$. The decoder $G$ reconstructs a data point $\hat{x} = G(z, s)$ from the bitstream $z$ and the sensitive attribute $s$. We model encoder and decoder as neural networks whose architecture varies with the type of data at hand.

The binarization layer controls explicitly the bit allowance of the learned representation and thus forces the encoder to strip redundancies – including sensitive attributes. Binarization is a two step process: (i) mapping the latent variable $e$ into $[0,1]^m$; (ii) converting real values into 0-1 bit. We achieve the first step by applying a neural network layer with an activation function $\tau = (\tanh(e) + 1)/2$. We achieve the second step by rounding $\tau$ to the closest integer 0 or 1. One issue with this approach is that the resulting binarizer $B$ is not differentiable with respect to $\tau$. To sidestep the issue, we follow Mentzer et al. (2018) or Theis et al. (2017) and rely on soft binarization during backward passes through the neural network. Formally, during a backward pass we replace $z$ by a soft-binary variable $\hat{z}$:

$$\hat{z} = \frac{\exp(-\sigma ||\tau - 1||_2^2)}{\exp(-\sigma ||\tau - 1||_2^2) + \exp(-\sigma ||\tau||_2^2)},$$

where $\sigma$ is an hyperparameter that controls the soft-binarization. During the forward pass, we use the binary variable $z$ instead of its soft-binary counterpart $\hat{z}$ to control the bitrate of the binary representation $Z$.

To estimate the entropy $H(z)$, we factorize the distribution $P(z)$ over $\{0,1\}^m$ by writing $z = (z_1, ... , z_m)$ (Mentzer et al. 2018) and by computing $P(z)$ as the product of conditional distributions:

$$P(z) = \prod_{i=1}^{m} p(z_i | z_{<i}), \quad (4)$$

where $z_{<i} = (z_1, z_2, ..., z_{i-1})$. The order of the bits $z_1, ..., z_m$ is arbitrary, but consistent across all data points. We model $P$ with a neural network $Q$ that predicts the value of each bit $z_i$ given the previous values $z_{i-1}, z_{i-2}, ..., z_1$. With the factorization (4), the entropy $H(z)$ is given by

$$H(z) = E_{z} \left[ \sum_{i=1}^{m} -\log(Q(z_i | z_{<i}))\right] - KL(P||Q), \quad (5)$$

where $CE(P,Q)$ is the cross entropy between $P$ and $Q$. Therefore, minimizing the cross-entropy loss of the neural network $Q$ minimizes an upper bound of the entropy of the code $z$. The encoder $F$ and the entropy estimator $Q$ cooperate. The lower the cross-entropy of $Q$ is, the lower is the estimate of the bit rate $H(z)$. Therefore, the encoder has incentives to make the bit stream easy to predict for the neural network $Q$. Designing a powerful predictor for the bit stream $z$ does not necessary complicate the loss landscape, unlike what could happen with adversarial methods (Berard et al. 2019). Since the prediction of $Q$ for the $i^{th}$ bit depends on the values of the previous bits $z_{i-1}, ..., z_1$, the factorization of

3In Pytorch, the binarizer returns $(z - \hat{z}).detach() + \hat{z}$. 

---

**Table 1:** Methods in unsupervised fair representation learning organized by whether the fairness properties of the learned representations is obtained by minimizing the mutual information between sensitive attributes $S$ and representations $Z$; or by minimizing the mutual information between data $X$ and representations $Z$; and whether $Z$ is modelled as a binary bit stream or is convolved with Gaussian noise.

| Methods | Fairness by controlling: | Examples |
|---------|---------------------------|----------|
| **Adversarial** | Minimizing auditor’s cross-entropy | \(x\) | Madras et al. (2018), Edwards and Storkey (2015), Creager et al. (2019) |
| **MMD** | Minimizing maximum mean discrepancy | \(x\) | Li, Swersky, and Zemel (2014), Louizos et al. (2015) |
| **$\beta$-VAE** | | \(\times\) | Higgins et al. (2017), This paper |
| **FBC** | | \(\times\) | This paper |
The objective of this experimental section is to demonstrate fair representations learning: $\beta$-VAE, Adv, MMD and VFAE.

(i) $\beta$-VAE (Higgins et al. 2017) solves the information bottleneck by variational inference and generates fair representations by adding Gaussian noise which upper-bounds the mutual information between $Z$ and $X$; (ii) MMD (Li, Swersky, and Zemel 2014) uses a deterministic auto-encoder and enforces fairness by minimizing the maximum mean discrepancy (Gretton et al. 2012) between the distribution of latent factors $Z$ conditioned on sensitive attributes $S$; (iii) VFAE (Louizos et al. 2015) extends $\beta$-VAE by adding a maximum mean discrepancy penalty; (iv) Adv (Edwards and Storkey 2015) uses a deterministic auto-encoder as for MMD, but enforces the fairness constraint by maximizing the cross-entropy of an adversarial auditor that predicts sensitive attributes $S$ from representations $Z$.

Although FBC shares the deterministic nature of Adv and MMD, it is more closely related to $\beta$-VAE, since $\beta$-VAE obtains fairness without explicit constraint on the mutual information of $I(Z, S)$. The main difference between our approach FBC and $\beta$-VAE is that FBC controls the entropy of a binary coding of the data, while $\beta$-VAE generates noisy representations and approximates the mutual information $I(Z, X)$ with the Kullback divergence between $Q(z|x)$ and a Gaussian prior $P(z)$. Note that the use of a vanilla $\beta$-VAE in a fairness context is novel: only its cousin VFAE with an additional MMD penalty has been proposed as a fair representation method.

Both FBC and $\beta$-VAE attempt to obtain fairness by controlling $I(Z, X)$. However, $\beta$-VAE assumes further that the prior distribution of the representation is an isotropic Gaussian. FBC does not require such a strong assumption and could still work even if the data is not generated from a factorized distribution. $\beta$-VAE is meant to compress and factorize. The main result from this paper is that compression is sufficient to learn fair representations and thus, disentanglement might be too restrictive. For problems where factorization could be hard to achieve in an unsupervised setting (Locatello et al. 2018), we would expect FBC to outperform $\beta$-VAE.

### 4.2 Experimental Protocol

The overall experimental procedure consists of:

(i) Training an encoder-decoder architecture $(F, B, G)$ along with an estimator of the code entropy $Q$;

(ii) Freezing its parameters;

(iii) Training an auditing network $Aud : Z \rightarrow S$ that predicts sensitive attributes from $Z$;

(iv) Training a task network $T : Z \rightarrow \mathcal{Y}$ that predicts a task label $Y$ from $Z$.

The encoder-decoder does not access the task labels during training: our representation learning approach is unsupervised with respect to downstream task labels. Datasets are split into a training set used to trained the encoder-decoder architecture; two test sets, one to train both task and auditing networks on samples not seen by the encoder-decoder; one to evaluate their respective performances.

**Pareto fronts.** To compare systematically performances across methods, we rely on Pareto fronts that estimates the maximum information that can be attained by a method for a given level of fairness. We approximate information content as the accuracy $A_y$ of the task network $T$ when predicting the downstream label $Y$. The larger $A_y$, the more useful is the learned representation for downstream task labels.
In the general case, $T$ parity that can be obtained by a downstream task classifier $D$ than binary ($\Delta$ chance level, the more independent $Z$ the accuracy $A$. Oquab (2016)). If the sensitive classes are furthermore balanced, the approximation is an upper bound to the true bit-rate (in nats) of $Z$. We estimate the distortion generated by the encoder-decoder procedure as the $l_2$ loss between reconstructed data $\hat{X} = G(F(X))$ and observed data $X$.

**Robustness to Fairness Metrics.** The fair information bottleneck (1) aims at controlling the flow of information between $Z$ and $S$. McNamara, Ong, and Williamson [2017] show that minimizing $I(Z, S)$ minimizes an upper bound of the demographic disparity $\Delta(T)$ of a task network $T$ that predicts a binary task label $Y$ from $Z$, where demographic parity $\Delta(T)$ is defined as

$$\Delta(T) = \sum_{s \in S} |P(T(x) = 1|S = s) - P(T(x) = 1|S \neq s)|.$$  

Moreover, the fair information bottleneck (1) is solved without a prior knowledge of specific downstream task labels $Y$. Therefore, (1) is not designed to control for fairness criteria that rely on labels $Y$ (e.g. equality of odds or opportunities, Hardt et al. [2016]) or on a specific classifier (e.g. individual fairness, Dwork et al. [2012]), unless downstream task labels are orthogonal to sensitive attributes conditional on features $X: Y \perp S|X$. In practice, we explore whether empirically FBC can generate representations that exhibit for a given task network $T$, low differences in false positive rates $\Delta FPR(T)$ with

$$\Delta FPR(T) \triangleq \sum_{s \in S} |P(T(x) = 1|Y = 0, S = s) - P(T(x) = 1|Y = 0, S \neq s)|.$$

**Rate distortion curves.** To demonstrate further our theoretical insights from section [2] we study both rate-distortion and rate-fairness curves of compressing methods FBC and $\beta$-VAE.

The rate-distortion function $RD(D)$ of an encoder-decoder is measured as the minimum bitrate (in nats) necessary for the distortion $E_{x,z,s}[-\log(p(X|Z,S)]$ to be less than $D$ (Tishby, Pereira, and Bialek [2000]).

$$RD(D) = \min I(Z, X) \text{ s.t. } E_{x,z,s}[-\log(p(X|Z,S)] \leq D.$$  

(6)

We introduce a new concept, rate-fairness function $RF(\Delta)$, and define it as the maximum bit rate allowed for the accuracy $A_s$ of the auditing classifier to remain less than $\Delta$:

$$RF(\Delta) = \max I(Z, X) \text{ s.t. } A_s \leq \Delta.$$  

(7)

The rate-fairness function captures the maximum information $Z$ can contain while keeping $A_s$ under a given threshold. To obtain both rate-distortion and rate-fairness curves for either our binary compression – FBC – or variational – $\beta$-VAE and VFAE – approaches, we vary the value of the parameter $\beta$ controlling the rate-distortion trade-off and for each value of $\beta$, we train the model 50 times with different seeds. For our binary compression method, FBC, the bit rate is approximated by the cross-entropy of the entropy estimator $Q$ in (5); for variational-based methods, the bit rate is approximated by the Kullback divergence between $Q(z|x)$ and a Gaussian prior. In both cases, the approximation is an upper bound to the true bit-rate (in nats) of $Z$. We estimate the distortion generated by the encoder-decoder procedure as the $l_2$ loss between reconstructed data $\hat{X} = G(F(X))$ and observed data $X$.
4.3 Datasets

First, we apply our experimental protocol to a synthetic dataset – DSprites Unfair – that contains 64 by 64 black and white images of various shapes (heart, square, circle). Images in the DSprites dataset are constructed from 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 modify the sampling to generate a source of potential unfairness and use as sensitive attribute a variable that encodes the quadrant of the circle the orientation angle belongs to.

Then, we extend our experimental protocol to three benchmark datasets in fair machine learning: Adults, Compas and Heritage. 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. Sensitive attributes is made of 10 categories that intersect gender and race to which individuals self-identify to. The downstream task label \(Y\) correspond to whether an individual earns more than 50K per year.

The Compas data contains 7K individuals with information related to their criminal history, misdemeanors, gender, age and race. Sensitive attributes intersect self-reported race and gender and result in four categories. The downstream task label \(Y\) assesses whether an individual presents a high risk of recidivism.

The Health Heritage dataset contains 220K individuals with 66 features related to age, clinical diagnoses and procedure, lab results, drug prescriptions and claims payment aggregated over 3 years. Sensitive attributes are 18 categories that intersect the gender which individuals self-identify to and their reported age. The downstream task label \(Y\) relates to whether an individual has a positive Charlson comorbidity Index.

5 Results and Discussion

5.1 Pareto Fronts

Figure 2 shows the Pareto fronts across five comparative methods for the DSprites and real-world datasets, respectively. Across all dataset, the higher and more leftward the Pareto front, the higher is the task accuracy \(A_y\) for a given auditor accuracy \(A_s\) and the better is the accuracy-fairness trade-off. From these Pareto fronts, we can draw three conclusions.

---

https://github.com/deepmind/dsprites-dataset/
https://archive.ics.uci.edu/ml/datasets/adult
https://github.com/propublica/compas-analysis/
https://foreverdata.org/1015/index.html
First, on all datasets, controlling for the mutual information between $Z$ and $X$ – as in FBC and $\beta - VAE$ – is sufficient to reduce the accuracy $A_s$ of the auditor $Aud$. This result is consistent with our theoretical observation that minimizing proxies for the information rate $I(Z, X)$ is sufficient to minimize $I(Z, S)$, provided that a side-channel provides the sensitive attributes $S$ to the decoder.

Second, in the $(A_s, A_y)$- plan, our method, FBC achieves either similar (Adults, Heritage) or better (DSprites, Compas) accuracy-fairness trade-off than the variational method $\beta - VAE$ that controls $I(Z, X)$ by adding noise to the information channel between $Z$ and $X$. Across all experiments, the Pareto fronts obtained from FBC are at least as upward and leftward as for $\beta - VAE$. This is consistent with our intuition that FBC may outperform $\beta - VAE$ in situations where disentanglement of the data into factorized representation is difficult (see (Locatello et al., 2018) for DSprites).

Third, FBC is a method that appears to be more consistently state-of-the-art in terms of performances compared to existing methods. FBC offers a better accuracy-fairness trade-off for Compas and DSprites than MMD, VFAE and Adv and is competitive for Adults and Heritage. This is true although Adv, VFAE and MMD control directly the mutual information between $Z$ and $S$, while FBC controls only $I(Z, X)$. The adversarial methods do not manage to generate representations with low $A_s$ for the DSprites dataset, possibly because in this higher dimensional problem, the optimization gets stuck in local minima where the adversary has no predictive power, regardless of the encoded representation.

5.2 Rate-distortion and Rate-fairness

Figure 3 confirms that for FBC, a lower bit rate estimated by the cross entropy $CE(p, q)$ corresponds to a lower accuracy for the auditing classifier $Aud$. Both rate-distortion $(R, D)$ and rate-fairness $(R, \Delta)$ curves show the same monotonic behavior: as distortion moves up along the rate-distortion curves, lack of fairness as measured by $A_s$ moves down. However, for real-word datasets, particularly for Adults and Comps, we observe more variance in the auditor accuracy’s $A_s$ given a representation bit rate. We attribute this higher variance to a smaller sample size – 617 for Comps and 3,256 for Adult on the test set.

Figure 4 shows that controlling for the level of compression by increasing the value of $\beta$ in $SNE$ allows moving smoothly along the rate-fairness curve. This is true whether the mutual information $I(Z, X)$ between data and representation is controlled by the bitstream entropy as in FBC (Figure 4) or by adding a noisy channel as in $\beta - VAE$ (see results in appendix). However, binary compression allows a tighter control of the fairness of the representation $Z$ than variational-based methods since in Figure 4, for a given auditor’s accuracy $A_s$, FBC allows the downstream classifier to achieve a higher accuracy $A_y$ while predicting $Y$ from $Z$.

5.3 Other Fairness Metrics

Figure 5 extends the pareto fronts of Figure 2 to additional fairness criteria. It plots the median accuracy obtained by task network $T$ against its differences in false positive rates $\Delta FPR$ and its demographic disparity $\Delta$.

First, all the methods tested – Adv, $\beta - VAE$ and FBC – generate an accuracy/fairness trade-off by reducing differences of false positive rates and demographic disparity at the cost of a lower downstream accuracy. Figure 5 illustrates a fairness transfer, where general purpose fair representations can offer some guarantees against some fairness criteria that the auto-encoder is not trained to minimize. This transfer is all the more remarkable for differences in false positive rates that rely on downstream task labels $Y$ that were not accessed by the auto-encoder during its training.

Second, for a given value of $\Delta FPR$ or $\Delta$, FBC reaches higher task accuracy $A_y$ than $\beta - VAE$ and is competitive with Adv for low values of $\Delta FPR$ and $\Delta$.

5.4 Representation Embeddings

Figure 6 shows the $t-SNE$ visualizations (Maaten and Hinton (2008)) of the representations generated by FBC for different values of the parameter $\beta$ that controls the rate-distortion trade-off in (3) for the Adults dataset. Without control of the representation bit rate $- \beta = 0$ – the $t-SNE$ plot show a cluster of Females that are isolated from males and thus, are easily detected by an auditor that predicts $S$ from $Z$.

With enough compression $- \beta = 0.35$ – the representation not only looks more parsimonious, but also does not separate Females from Males as much as without compression ($\beta = 0$). In the embeddings space, Females plots are either within clusters of Males or on the edges of these clusters. Moreover, the $t-SNE$ visualizations separate individuals by income level regardless of the compression level, which confirms that the representations generated by FBC are useful for classification tasks that predict income level from $Z$.

To quantitatively assess the local homogeneity of the sensitive attribute in the embedding space (Figure 6 top), we compute the average distance of females to their top-10 male neighbors and normalize it by the average distance between all individuals. We find that our homogeneity measure decreases by 30% when compressing the data (from left to right plot). But, a similar measure of homogeneity for outcomes (bottom row) decreases only by 8%. This result confirms the visual perception that compression decreases the local homogeneity of sensitive attributes more than the homogeneity of downstream task labels.

6 Conclusion

This paper introduces a new method – Fairness by Binary Compression (FBC) – to map data into a latent space, while guaranteeing that the latent variables are independent of sensitive attributes. Our method is motivated by the observation that in an information bottleneck framework, controlling for the mutual information between representation and data is sufficient to remove unwanted factors, provided that these unwanted factors are direct inputs to the decoder.

Our empirical findings confirm our theoretical intuition: FBC offers a state-of-the-art accuracy-fairness trade-off
Figure 5: Differences in false positive rates and demographic disparity of downstream task networks. This shows pareto fronts for Adults and Compas as in 2 but using $\Delta (\ref{eq:fairness_criteria}) \Delta FPR (\ref{eq:fairness_criteria})$ as a fairness criteria. Shaded areas show the area between the 25 – th and 75 – th quantiles of the pareto front.

Figure 6: Adults – t-SNE visualizations colored with gender ($S$) and income level ($Y$) of the representations obtained by FBC for different values of the parameter $\beta$ controlling the compression rate of FBC.

across four benchmark datasets. Moreover, we observe that encoding the representation into a binary stream allows a tighter control of the fairness-accuracy trade-off than limiting the information channel capacity by adding noise. Our results suggest further research into encoder-decoder whose architecture allows a tighter control of the representation’s bit rate and thus, of its fairness.

Acknowledgments

This work is supported by the National Science Foundation grant No. 1937950.

References

Achille, A.; and Soatto, S. 2018. Emergence of Invariance and Disentanglement in Deep Representations. *Journal of Machine Learning Research* 19(50): 1–34. URL [http://jmlr.org/papers/v19/17-646.html](http://jmlr.org/papers/v19/17-646.html).

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

Agustsson, E.; Mentzer, F.; Tschannen, M.; Cavigelli, L.; Timofte, R.; Benini, L.; and Gool, L. V. 2017. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, 1141–1151.

Berard, H.; Gidel, G.; Almahairi, A.; Vincent, P.; and Lacoste-Julien, S. 2019. A closer look at the optimization landscapes of generative adversarial networks. *arXiv preprint arXiv:1906.04848*.

Braithwaite, D. T.; and Kleijn, W. B. 2018. Bounded information rate variational autoencoders. *arXiv preprint arXiv:1807.07306*.

Buolamwini, J.; and Gebru, T. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 77–91. New York, NY, USA: PMLR. URL [http://proceedings.mlr.press/v81/buolamwini18a.html](http://proceedings.mlr.press/v81/buolamwini18a.html).

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

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M. A.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*.

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*, 214–226. ACM.

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
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), 570–575. IEEE.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In International Conference on Machine Learning, 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, 335–340.