SoFaiR: Single Shot Fair Representation Learning

Xavier Gitiaux and **Huzefa Rangwala**
George Mason University
{xgitiaux, rangwala}@gmu.edu

Abstract

To avoid discriminatory uses of their data, organizations can learn to map them into a representation that filters out information related to sensitive attributes. However, all existing methods in fair representation learning generate a fairness-information trade-off. To achieve different points on the fairness-information plane, one must train different models. In this paper, we first demonstrate that fairness-information trade-offs are fully characterized by rate-distortion trade-offs. Then, we use this key result and propose SoFaiR, a single shot fair representation learning method that generates with one trained model many points on the fairness-information plane. Besides its computational saving, our single-shot approach is, to the extent of our knowledge, the first fair representation learning method that explains what information is affected by changes in the fairness / distortion properties of the representation. Empirically, we find on three datasets that SoFaiR achieves similar fairness-information trade-offs as its multi-shot counterparts.

1 Introduction

Machine learning algorithms increasingly support decision-making systems in contexts where outcomes have long-term implications on the subject's well-being. A growing body of evidence find that algorithms can either replicate or exacerbate existing social biases against some demographic groups. These evidence span many domains, including recidivism risk assessment [ProPublica, 2016], face recognition [Buolamwini and Gebru, 2018], education data mining [Gardner *et al.*, 2019], and medical diagnosis [Pfohl *et al.*, 2019].

As a result, organizations that collect data are increasingly scrutinized for the potentially discriminatory use of a data by downstream applications. A flexible solution to the data-science pipeline is to control unfair uses of a data before its ingestion by a machine algorithm. Fair representation learning [Zemel *et al.*, 2013] follows this paradigm. It is a data pre-processing method that encodes the data into a representation or code Z, while removing its correlations with sensitive attributes S.

Current approaches in fair representation learning [Zemel *et al.*, 2013; Madras *et al.*, 2018; Gitiaux and Rangwala, 2021a; Creager *et al.*, 2019] generate a fairness-information trade-off and are inflexible with respect to their fairness-information trade-off, which is set at training time. This limits the deployment of fair representation learning approaches. For example, in medical applications, at test time, a user may need to adjust the content of the representation depending on whether gender is an appropriate feature for the downstream task at play. On one hand, for a downstream application that predicts cardiovascular risk, gender is an important/appropriate feature that should be part of the representation of the data. On the other hand, for a downstream application that predicts payment of medical bills, gender should be irrelevant to the outcome and thus, filtered out from the representation. With existing methods in fair representation learning, the user would have to re-train a fair encoder-decoder to meet each request. At issue are computational costs and lack of consistency between released representations, since the user cannot explain what changes occur between each data product it releases.

This paper introduces SoFaiR, **S**ingle Sh**o**t **Fai**r **R**epresentation, a method to generate a unfairness-distortion curve with *one single trained model*. We first show that we can derive unfairness-distortion curves from rate-distortion curves. We can control for the mutual information $I(Z, S)$ between representation and sensitive attribute by encoding X into a bitstream and by controlling for its entropy. We then construct a gated architecture that masks partially the bitstream conditional on the value of the Lagrangian multiplier in the rate-distortion optimization problem. The mask adapts to the fairness-information trade-off targeted by the user who can explore at test time the entire unfairness-distortion curve by increasingly umasking bits. For example, in the case of a downstream medical application for which gender is sensitive and needs to be filtered out, the user sets at test time the Lagrangian multiplier to its largest value, which lowers the resolution of the representation and in a binary basis, masks the rightmost tail of the bit stream.

Besides saving on computational costs, SoFaiR allows users to interpret what type of information is affected by movement along unfairness-distortion curves. Moving upward unmasks bits in the tail of the bitstream and thus, increases the resolution of the representation encoded in a binary basis. By correlating these unmasked bits with data features, the practitioner has at hand a simple method to explore what information related to the features is added to the representation as its fairness properties degrade.

Empirically, we demonstrate on three datasets that at cost constant with the number of points on the curve, SoFaiR con-

structs unfairness-distortion curves that are comparable to the ones produced by existing multi-shot approaches whose cost increases linearly with the number of points. On the benchmark Adults dataset, we find that increasingly removing information related to gender degrades first how the representation encodes working hours; then, relationship status and type of professional occupations; finally, marital status.

Our contributions are as follows: (i) we formalize fairness-information trade-offs in unsupervised fair representation learning with unfairness-distortion curves and show a tractable connection with rate-distortion curves; (ii) we propose a single shot fair representation learning method to control fairness-information trade-off at test time, while training a single model; and, (iii) we offer a method to interpret how improving or degrading the fairness properties of the resulting representation affects the type of information it encodes.

Proofs of theoretical results are in the appendix. Additional experimental details and results are in the supplementary file[1]. The code publicly available here[2].

Related Work. A growing body of machine learning literature explores how algorithms can adversely impact some demographic groups (e.g individuals self-identified as Female or African-American) (see [Chouldechova and Roth, 2018] for a review). This paper is more closely related to methods that transform the data into a fair representation. Most of the current literature focus on supervised techniques that tailor the representations to a specific downstream task (e.g [Madras *et al.*, 2018; Edwards and Storkey, 2016; Moyer *et al.*, 2018; Gupta *et al.*, 2021; Jaiswal *et al.*, 2020]). However, practical implementations of fair representation learning would occur in unsupervised setting where organizations cannot anticipate all downstream uses of a data. This paper contributes to unsupervised fair representation (e.g [Gitiaux and Rangwala, 2021b]) by (i) formalizing fairness-information trade-off in a distortion-rate phase diagram, which extends compression-based approaches (e.g [Gitiaux and Rangwala, 2021a]); and (ii), proposing an adaptive technique that allows a single trained model to output as many points as desired on a unfairness-distortion curve.

The implementation of SoFaiR relates to approaches in rate-distortion that learn adaptive encoder and vary the compression rate at test time (e.g. [Theis *et al.*, 2017; Choi *et al.*, 2019]. We borrow soft-quantization techniques and entropy coding to solve the rate-distortion problem that can be derived from the fair representation learning objective. Our adaptive mask relates to the gain function in [Cui *et al.*, 2020] that selects channels depending on the targeted bit rate. We rely on successive refinement methods from information theory (e.g [Kostina and Tuncel, 2019]) that use a common encoder for all points on the unfairness-distortion curve and add new information by appending bits to a initially coarse representation. To our knowledge, we are the first contribution to implement a deep learning multi-resolution quantization and apply it to the problem of single shot fair representation learning.

[1]See Long version at https://arxiv.org/abs/2204.12556

[2]See https://github.com/Gitiauxx/SoFaiR

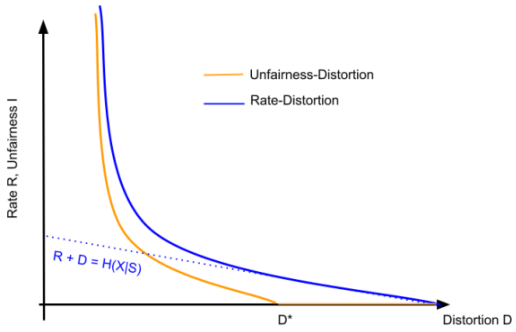

Figure 1: Unfairness-distortion curves $I(D)$ vs. rate-distortion curve $R(D)$. The unfairness distortion $I(D)$ can be deduced from the rate-distortion $R(D)$ curve by a downward shift equal to $D - H(X|S)$ if the distortion is less than D^*.

2 Problem Statement

2.1 Preliminaries

Consider a population of individuals represented by features $X \in \mathcal{X}$ and sensitive attributes in $S \in \mathcal{S} \subset \{0,1\}^{d_s}$, where $d_s \geq 1$ is the dimension of the sensitive attributes space.

The objective of unsupervised fair representation learning is to map features $X \in \mathcal{X}$ into a $d-$dimensional representation $Z \in \mathcal{Z}$ such that (i) Z maximizes the information related to X, but (ii) minimizes the information related to sensitive attributes S. We control for the fairness properties of the representation Z via its mutual information $I(Z,S)$ with S. $I(Z,S)$ is an upper bound to the demographic disparity of any classifier using Z as input [Gupta *et al.*, 2021]. We control for the information contained in Z by constraining a distortion $d(X,\{Z,S\})$ that measures how much information is lost when using a data reconstructed from Z and S instead of the original X. Therefore, fair representation learning is equivalent to solving the following unfairness-distortion problem

$$I(D) = \min_{f} I(Z,S) \text{ s.t. } d(X,\{Z,S\}) \leq D \qquad (1)$$

where $f : \mathcal{X} \to \mathcal{Z}$ is an encoder. The unfairness-distortion function $I(D)$ defines the minimum mutual information between Z and S a user can expect when encoding the data with a distortion less or equal to D. The unfairness-distortion problem (1) implies a fairness-information trade-off: lower values of the distortion constraint D degrade the fairness properties of Z by increasing $I(D)$. *The objective of this paper is given a data X to obtain the unfairness-distortion function $I(D)$ with a single encoder-decoder architecture.*

2.2 Unfairness Distortion Curves

Rate distortion theory characterizes the minimum average number of bits $R(D)$ used to represent X by a code Z while the expected distortion incurred to reconstruct X from the code is less than D. We show how to derive unfairness-distortion functions $I(D)$ from rate distortion functions $R(D)$.

Theorem 2.1. *Suppose that the distortion is given by* $d(X,\{Z,S\}) = E[-\log(p(x|z,s)]$. *Then, the unfairness*

distortion function $I(D)$ is equal to $R(D) + D - C$ if $\frac{\partial R}{\partial D} \leq -1$ and 0 otherwise. $C = H(X|S)$ is a constant that does not depend on D, but only on the data X. Moreover, $I(D)$ is a non-increasing convex function.

Phase diagram. Figure 1 shows a graphical interpretation of Theorem 2.1 in a (D, R) plane. (D^*, R^*) denotes the point on the rate-distortion curve where $\frac{\partial R}{\partial D} = -1$. For $D \leq D^*$, the rate distortion curve is above the line defined by $R + D = H(X|S)$ and that difference between $I(D)$ and $R(D)$ is $I(Z, S)$. For $D > D^*$, the rate-distortion curve is the line $R + D = H(X|S)$ and the unfairness-distortion curve is the horizontal axis. We call the regime $D^* \leq D \leq H(X|S)$ the fair-encoding limit where the distortion is less than its upper limit, but Z is independent of sensitive attribute S.

Information bottleneck. Theorem 2.1 implies that fairness-distortion trade-offs are fully characterized by rate-distortion trade-offs. A fundamental result in rate distortion theory ([Tishby, 1999]) shows that the rate-distortion function is given by the information bottleneck

$$R(D) = \min_f I(X, Z) \text{ s.t } d(X, \{Z, S\}) \leq D. \qquad (2)$$

By solving this information bottleneck with $d(X, \{Z, S\}) = H(X|Z, S)$ and invoking Theorem 2.1, we can recover the unfairness-distortion $I(D)$. [Gitiaux and Rangwala, 2021a] provide an intuition for this result. Controlling for the mutual information $I(Z, X)$ allows to control for $I(Z, S)$ because an encoder would not waste code length to represent information related to sensitive attributes, since sensitive attributes are provided directly as an input to the decoder. We can write the information bottleneck in its Lagrangian form as

$$\min_f \beta I(Z, X) + E[-\log p(x|z, s)] \qquad (3)$$

The coefficient β relates to the inverse of the slope of the rate-distortion curve: $\frac{\partial R}{\partial D} = -1/\beta$. Each value of β generates a different point along the rate-distortion curve and thus, by Theorem 2.1 a different point along the unfairness-distortion curve. Higher values of β lead to representations with lower bit rate and lower mutual information with S. To explore a unfairness-distortion curve, existing multi-shot strategies are prohibitively expensive as they learn a new encoder f for each value of β. Moreover, they cannot interpret how changes in β affect the representation generated by the encoder.

3 Method: Single-Shot Unfairness-Distortion Curves.

We propose a single-shot method, SoFaiR, to generate with one model as many points as desired on the unfairness-distortion curve. An encoder $f : \mathcal{X} \to \{0, 1\}^{d \times r}$ common to all values of β encodes the data into a d dimensional latent variable $e \in [0, 1]^d$. We quantize each dimension e_j of the $d-$dimensional latent variable with a resolution $r_j(\beta)$: we transform e_j into a quantized representation $z_j(\beta) = [e * r(\beta)]/r(\beta)$, where [.] denotes the rounding-up operation and $r(.)$ is a decreasing function of β.

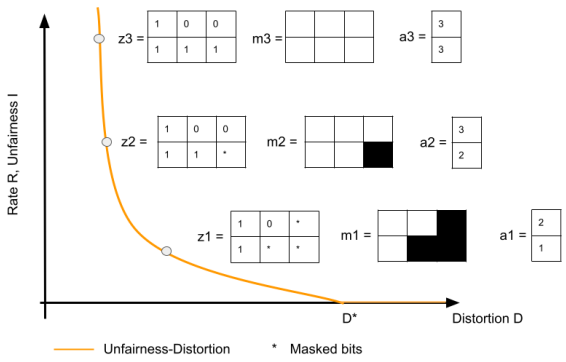

Figure 2: SoFaiR generates interpretable shifts along the unfairness-distortion curve. For a point $z1$, SoFair learns a mask $m1$ that hides bits on the tails of each dimension of the representation. By relaxing the mask to first $m2$ then $m3$, the number of bits used to represent the data increases from $a1$ to $a2$ and then $a3$; and, the representation moves to $z2$ then $z3$, which reduces the distortion at the expenses of degraded fairness properties. $z1$, $z2$ and $z3$ only differ by their masked bits (black squares).

3.1 Interpretability

To maintain an interpretable relation between $z(\beta)$ and $z(\beta')$ for $\beta' < \beta$, we write $r_j(\beta) = 2^{a_j(\beta)}$, where $a_j(.)$ is a decreasing function of β for $j = 1, ..., d$. Each dimension $z_j(\beta)$ of the quantized representation is then encoded into $a_j(\beta)$ bits. Moreover, for $\beta' < \beta$, each dimension j of the representation $z(\beta')$ is made of the same $a_j(\beta)$ bits as $z_j(\beta)$, followed by $a_j(\beta') - a_j(\beta)$ additional bits. Each dimension $z_j(\beta)$ of the quantized representation is encoded into $a_j(\beta)$ bits $b_{j,1}, b_{j,2}, ..., b_{j,a_j(\beta)}$, where $b_{j,l} \in \{0, 1\}$ for $l = 1, ..., a_j(\beta)$. For $\beta' < \beta$ and for $j = 1, ..., d$, we have

$$z_j(\beta') = z_j(\beta) + \sum_{l=a_j(\beta)}^{a_j(\beta')} b_{j,l} 2^{-l}.$$

Therefore, we have a tractable and interpretable relation between $z_j(\beta')$ and $z_j(\beta)$. This construction allows relaxing fairness constraints and decreasing distortion by unmasking additional bits for each dimension of the representation. Figure 2 shows an example for a 2-dimensional representation. A user who has released $z1$ with high distortion and low mutual information $I(Z, S)$ reduces distortion at the cost of fairness by unmasking one bit for the first dimension and two bits for the second and by generating $z2$.

3.2 Quantization

We assign a maximum number of bits $A > 0$ to encode each dimension of the representation. We apply a function h_e to map the $d-$dimensional latent variable e into $[0, 1]^{d \times A}$ and then, apply a rounding-up operator $[h_e(e)]$ to generate a $d \times A$ matrix, each row encoding a dimension of the representation with A bits (see Figure 2 with $A = 3$). For each dimension j, we implement $a_j(.)$ by applying a function h_a to map e into a $d-$dimensional vector of $\mathbb{R}+^d$ and by computing $a_j(\beta) = A\,[1 - \tanh(h_a(e)_j \beta))]$.

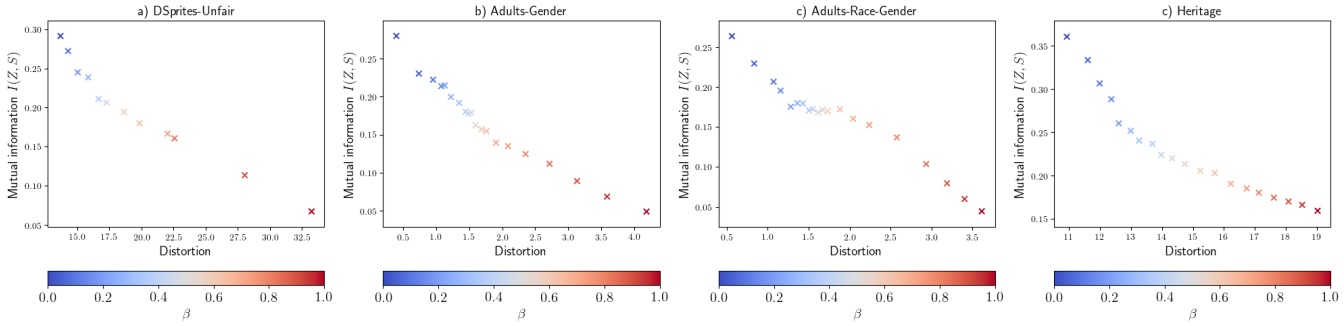

Figure 3: Unfairness-Distortion curves for a) DSprites, b) Adults-Gender, c) Adults-Race-Gender(left) and d) Heritage.

For each value of β and each row of the matrix $[h_e(e)]$, we mask all the entries in position $l > a_j(\beta)$: for each row j and each column l, we compute a soft mask $m_{j,l}(\beta) = \sigma\left(a_j(\beta) - l\right)$ where σ denotes a sigmoid activation; and then, we apply a rounding operator $[m_{j,l}(\beta)]$ to our soft mask.

For example, suppose that we encode in at most $A = 8$ bits the embedding value $e = 0.7$ and that $h(e) = e$. For $\beta = 0$, we use all the bits ($a(0) = 8$) and $z = 0.699$; for $\beta = 0.5$, $a = 8(1 - \tanh((0.5)(0.7))) = 5.3$ and we use only 5 bits with $z = 0.6875$.

The binarization caused by the rounding operation [.] is not differentiable. We follow [Theis *et al.*, 2017] and use a gradient-through approach that replaces [.] by the identity during the backward pass of back-propagation, while keeping the rounding operation during the forward pass.

3.3 Entropy Estimation

In our implementation, encoding and quantization are deterministic and Z is completely determined by X: $H(Z|X) = 0$ and $I(Z, X) = H(Z)$. To estimate the entropy of the representation Z, we use an auto-regressive factorization and write the discrete distribution $P(z|\beta)$ as $P(z|\beta) = \prod_{j=1}^{d} P(z_j|z_{.<j}, \beta)$, where the order of the dimension j is arbitrary and $z_{.<j}$ denotes the dimension between 1 and $j-1$.

We approximate the discrete distribution $p(z_j|z_{.<j}, \beta)$ by a continuous distribution $q(z_j|z_{.<j}, \beta)$ such that the probability mass of q on the interval $[z_j - 1/2^{a_j(\beta)}, z_j + 1/2^{a_j(\beta)}]$ is equal to $p(z_j|z_{.<j}, \beta)$. We can show then that $H(z|\beta)$ is bounded above by $\sum_{j=1}^{d} E_{z\sim p(z)} \log \int_{-1/2^{a_j(\beta)}}^{1/2^{a_j(\beta)}} q(z_j + u|z_{.<j}, \beta)du$ (see Appendix). We follow [Salimans *et al.*, 2017] and for each $j = 1, ..., d$, we model $q(.|z_{.<j}, \beta)$ as a mixture of K logistic distributions with means $\mu_{j,k}(\beta)$, scales $\gamma_{j,k}(\beta)$ and mixtures probability $\pi_{j,k}(\beta)$, which allows a tractable formulation of our upper-bound (see Appendix). The resulting adaptive information bottleckneck (3) is:

$$\min_{g,f,q,\mu,\gamma,\pi} E[-\log p(x|g(z, s, \beta)) + \beta H(z|\beta)], \qquad (4)$$

where g is a decoder that reconstructs the data x from z, s and β. The expectation is taken over the data x and values of β uniformly drawn in $[0, 1]$.

4 Experiments

We design our experiments to answer the following research questions: (RQ1) Does SoFaiR generate in a single-shot unfairness-distortion curves comparable to the ones generated by multi-shot models? (RQ2) Do representations learned by SoFaiR offer to downstream tasks a fairness-accuracy trade-off on par with state-of-the-art multi-shots techniques in unsupervised fair representation learning? (RQ3) What information is present in the additional bits that are unmasked as we move up the unfairness-distortion curve? Architecture details and hyperparameter values are in the supplementary file.

4.1 Datasets

We validate our single-shot approach with three benchmark datasets: **DSprite-Unfair**, **Adults** and **Heritage**.

DSprite Unfair is a variant of the DSprites data and contains 64 by 64 black and white images of various shapes (heart, square, circle). We modify the sampling to generate a source of potential unfairness and use as sensitive attribute a variable that encodes the orientation of the shapes.

The **Adults** dataset contains $49K$ individuals with information on professional occupation, education attainment, capital gains, hours worked, race and marital status. We consider as sensitive attribute, gender in **Adults-Gender**; and, gender and race in **Adults-Gender-Race**.

The **Health Heritage** dataset contains $95K$ individuals with 65 features related to clinical diagnoses and procedures, lab results, drug prescriptions and claims payment aggregated over 3 years (2011-2013). We define as sensitive attributes an intersection variable of gender and age.

4.2 Unfairness-Distortion Curves

To plot unfairness-distortion curves, we estimate the distortion as the l_2- loss between reconstructed and observed data, which is equal to $E_{x,z,s}[-\log p(x|z, s)]$ (up to a constant) if the distribution of $p(X|Z, S)$ X is an isotropic Gaussian. We also approximate the mutual information $I(Z, S)$ with an adversarial lower bound (see appendix):

$$I(Z, S) \geq H(S) - \min_{c} E_{s,z}[-\log c(s|z)], \qquad (5)$$

where c is an auditing classifier that predicts S from Z. Unlike adversarial methods (e.g. [Edwards and Storkey, 2016]), we do not use this bound for training our encoder-decoder, but

Dataset	Model	AUFDC (↓)	Average per step (ms) CPU / GPU	Total time (10^6 ms): CPU/GPU (↓) 4 points	8 points	16 points
DSprites-UnfaiR	SoFaiR	0.21	79 ± 1.2 / 55 ± 0.2	**18.5/13.0**	**18.5/13.0**	**18.5/13.0**
	SoFaiR-NOS	0.25	78 ± 1.1 / 54 ± 0.3	**18.4/13.1**	**18.4/13.1**	**18.4/13.1**
	MSFaiR	**0.14**	76 ± 3.2 / 55 ± 0.3	71.4/52.1	142.9/104.2	285.8/208.0
Adults-Gender	SoFaiR	**0.32**	$91 \pm 3.3/6 \pm 0.0$	**2.3/0.1**	**2.3/0.1**	**2.3/0.1**
	SoFaiR-NOS	0.58	$91 \pm 4.3/6 \pm 0.0$	**2.3/0.1**	**2.3/0.1**	**2.3/0.1**
	MSFaiR	0.35	$92 \pm 1.0/6 \pm 0.0$	9.4/0.6	18.9/1.1	37.7/2.3
Adults-Gender-Race	SoFaiR	**0.30**	$92 \pm 4.3/6 \pm 0.0$	**2.4/0.1**	**2.4/0.1**	**2.4/0.1**
	SoFaiR-NOS	0.53	$92 \pm 4.0/6 \pm 0.0$	**2.4/0.1**	**2.4/0.1**	**2.4/0.1**
	MSFaiR	0.36	$90 \pm 4.0/6 \pm 0.0$	9.1/0.6	18.3/1.1	36.6/2.3
Heritage	SoFaiR	0.62	$125 \pm 3.0/8.6 \pm 1.6$	**3.7/0.3**	**3.7/0.3**	**3.7/0.3**
	SoFaiR-NOS	0.73	$123 \pm 2.5/10 \pm 0.3$	**3.7/0.3**	**3.7/0.3**	**3.7/0.3**
	MSFaiR	**0.56**	$123 \pm 3.1/10 \pm 0.8$	14.7/1.2	29.4/2.3	58.7/4.8

Table 1: Area under the unfairness-distortion curve and computational costs of single-shot (SoFaiR) versus multi-shot (MSFaiR) fair representation learning methods. Lower (↓) is better. This shows that SoFaiR provides unfairness-distortion curves with similar AUFDC as MSFaiR, but at much lower computational costs.

only for post mortem evaluation of the unfairness-distortion trade-off generated by SoFaiR. In practice, we train a set of 5 fully connected neural networks $c : \mathcal{Z} \to \mathcal{S}$ and use their average cross-entropy to estimate the right hand side of (5).

4.3 Area Under Unfairness-Distortion Curves

To quantitatively compare the unfairness-distortion curves of competing approaches, we introduce the area under unfair-distortion curve, AUFDC. A lower AUFDC means that a model achieve lower $I(Z, S)$ for a given level of distortion. To allow comparison across datasets, we normalize the value of AUFDC by the area of the rectangle $[0, D_{max}] \times [0, I_{max}]$, where D_{max} is the distortion obtained by generating random permutation of a representation and I_{max} is the value of the lower bound (5) when auditing raw data.

4.4 Comparative Methods

Methods. We compare SoFaiR with five fair representation methods: (i) **LATFR** (e.g [Madras *et al.*, 2018]) controls for $I(Z, S)$ by using the lower bound (5) ; (ii) **MaxEnt-ARL** [Roy and Boddeti, 2019] replaces the adversary's cross-entropy of LATFR with the entropy of the adversary's predictions; (iii) **CVIB** [Moyer *et al.*, 2018] replaces adversarial training with an information-theory upper bound of $I(Z, S)$; (iv)$\beta - VAE$ [Higgins *et al.*, 2016] solves the information bottleneck (3) by variational inference, which upper-bounds $I(Z, S)$, provided that the decoder uses the sensitive attribute as input [Gitiaux and Rangwala, 2021a]; (v) **MSFaiR** reproduces SoFaiR, but solves the rate-distortion problem (4) separately for different values of β. All methods have the same autoencoder architecture. Most methods are tailored to a specific downstream task. In our unsupervised setting, we repurpose them by replacing the cross-entropy of the downstream classification task with our measure of distortion $E[-\log p(x|Z, s)]$.

Pareto fronts. We construct Pareto fronts that compare the unfairness properties of the representation to the accuracy A_y of a downstream task classifier that predicts a downstream label Y from Z. Critically in our unsupervised setting, we do not provide the labels Y to encoder-decoders. To match existing benchmarks, we measure the unfairness properties of the representation with the average accuracy A_s of auditing classifiers that predict S from Z. The higher A_y for a given A_s, the better is the fair representation method.

5 Results

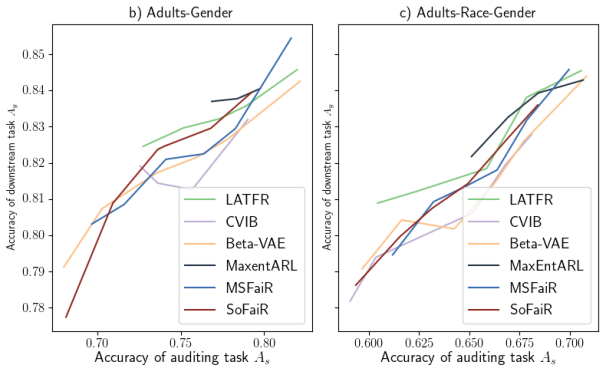

Figure 4: Pareto fronts for a) Adults-Gender (left), b) Adults-Race-Gender(right). The downstream task label is whether income is larger than 50K.

5.1 RQ1: Single Shot Fairness-Distortion Curves

Figure 3 shows SoFaiR's unfairness-distortion curves for DSprites (left), Adults-Gender (middle left), Adults-Gender-Race (middle right) and Heritage (right). By increasing at test time the value of β, the user can smoothly move down the unfairness-distortion curve: values of β close to zero lead to low distortion - high $I(Z, S)$ points; values of β close to one lead to higher distortion - low $I(Z, S)$ points. Figure 3 demonstrates that a solution to the adaptive bottleneck (4) allows one single model to capture different points on the unfairness-distortion curve. This result is consistent with Theorem 2.1 and illustrates that controlling for the bit rate of Z via its entropy $H(Z)$ is sufficient to control for $I(Z, S)$.

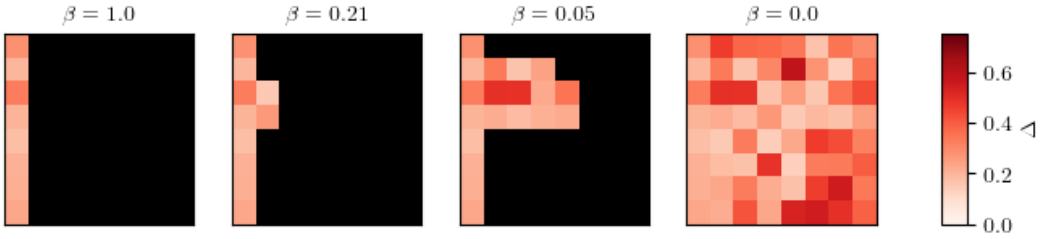

Figure 5: Unmasked bits for different values of the fairness coefficient β for the Adults-Gender-Race dataset. Each row is a dimension of Z. Each colored square is an unmasked bit. Black squares represent masked bits. Darker bits exhibit higher bit demographic disparity $\Delta(b)$. As β decreases, SoFaiR unmasks more bits for each dimension of Z. And, bits with higher disparity are more likely to be the last unmasked.

Ablation study. AUFDC scores in Table 1 show that SoFaiR is competitive with its multi-shot counterpart: SoFaiR outperforms MSFaiR for Adults-Gender and Adults-Gender-Race (lower AUFDC), but is slightly outperformed for Heritage and DSprites-Unfair (higher AUFDC). On the other hand, SoFaiR unambigously outperforms SoFaiR-NOS, a model similar to SoFaiR but with a decoder that does not use the sensitive attribute S as side-channel. The relation between unfairness-distortion and rate-distortion curves in Theorem 2.1 is tractable only if we use $E[-\log(p(x|z,s)]$ as a measure of distortion and does not hold if we use $E[-\log(p(x|z)]$ instead and the decoder does not receive S as side channel.

Computational costs. Table 1 compares the computational costs of SoFaiR and MSFaiR. We average the cpu and gpu times of a training step over 10 profiling cycles and the number of training epochs. We perform the experiment on a AMD Ryzen Threadripper 2950X 16-Core Processor CPU and a NVIDIA GV102 GPU. The average computing cost of a training step is similar for SoFaiR and MSFaiR since both methods rely on similar architecture. However, SoFaiR's computational costs remain constant as the number of points on the unfairness-distortion curve increases, while MSFaiR's costs increase linearly. For example, 16 points for the DSprites-Unfair require about 137 hours of running time with MSFaiR and only 8 hours with SoFaiR.

5.2 RQ2: Pareto Fronts

In Figure 4, the larger the downstream classifier's accuracy A_y for a given value of the auditor's accuracy A_s, the better the Pareto front. First, SoFaiR and MSFaiR's Pareto fronts are either as good or better than the ones generated by $LATFR$, $CVIB$, $Maxent-ARL$ and $\beta-VAE$. Exceptions to this observations include Adults-Gender-Race for low values of A_s where $LATFR$ outperforms SoFaiR/MSFaiR. Rate distortion approaches are competitive, which confirms the tight connection between rate-distortion and unfairness-distortion as presented in Theorem 2.1. Both SoFaiR and MSFaiR offer more consistent performances than $LATFR$ or $Maxent-ARL$ whose representations keep leaking information related to S for Adults-Gender regardless of the constraints placed on the adversary. And, $\beta-VAE$ exhibits non-monotonic behavior for Adults-Gender. Second, Figure 4 shows that SoFaiR's Pareto fronts are similar to the ones offered by MSFaiR, its multi-shot counterpart. This result is consistent with AUFDC scores in Table 1. Pareto fronts for Heritage and DSPrites-Unfair are in the supplementary file.

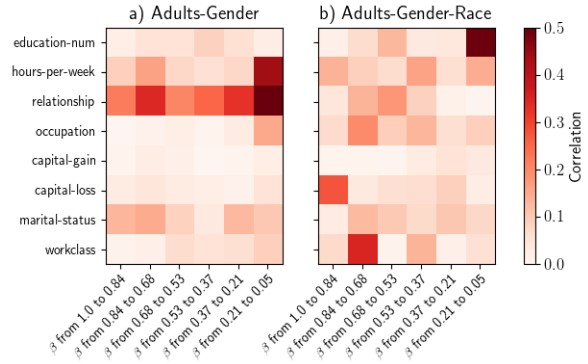

Figure 6: Additional information provided by refining the representation for Adults-Gender (left) and Adults-Gender-Race (right) dataset. This shows the correlation between data features and additional bits that SoFaiR unmasks when loosening the fairness constraint. Correlations are computed between the data features and the first principal component of newly unmasked bits. Each column corresponds to a decrease of β as labeled on the horizontal axis.

5.3 RQ3: Interpretability

Bit disparity. We measure the disparity of each bit b as $\Delta(b) = \max_{s\in\mathcal{S}} |P(b=1|S=s) - P(b=1|S\neq s)|$. Bit disparity is the demographic disparity of a classifier that returns 1 if $b=1$ and 0 otherwise. Moreover, we show in the supplementary file that $\max_b \Delta(b)$ is a lower bound of $I(Z,S)$: a large value of $\Delta(b)$ means that the presence of bit b in the bitstream will significantly degrade the fairness properties of Z. In Figure 5, loosening the fairness constraint at test time – decreasing β – unmasks more bits, while keeping the leftmost bits identical to ones obtained with higher values of β. SoFaiR degrades gracefully the fairness properties of the representation by increasing its resolution. Figure 5 also shows that for Adults-Gender-Race, bits with higher disparity Δ are less likely to be unmasked with stringent fairness constraints – high β – and are only active when more leakages related to sensitive attribute are tolerated – low β. Therefore, by forcing SoFaiR to generate many points on the unfairness-distortion curve, we obtain an information ordering that pushes to the tail of the bitstreams the bits the most

correlated with S. We observe a similar pattern with Adults-Gender (supplementary file).

Fairness and information loss. Unlike alternative methods in fair representation learning, SoFaiR offers a simple tool to interpret at test time what information is lost as the fairness constraint tightens. In Figure 6, we plot for Adults-Gender and Adults-Gender-Race how additional bits unmasked as β decreases correlate with data features. As we move up the unfairness-distortion curve for Adults-Gender, additional information first relates to marital status; then, occupation type, relationship status and hours-per-week. It means that for downstream tasks that predict marital status, a representation on the bottom right of the unfairness-distortion curve (high distortion, low $I(Z,S)$) is sufficient to achieve good accuracy. But, downstream tasks that need hours-per-week would find more difficult to obtain good accuracy without moving up the unfairness-distortion curves, i.e leaking additional information related to sensitive attribute S.

6 Conclusion

In this paper, we present SoFaiR, a single-shot fair representation learning method that allows with one trained model to explore at test time the fairness-information trade-offs of a representation of the data. Our implementation relies on a tight connection between rate-distortion and unfairness-distortion curves. SoFaiR is a step toward practical implementation of unsupervised fair representation learning approach, all the more as users can now explain what information is lost as the fairness properties of the representation improve.

A Appendix

A.1 Proof of Theorem 2.1

First, we show the following identity:

Lemma A.1. $I(Z,S) = I(Z,X) + H(X|Z,S) - H(X|S)$.

Proof. The proof of Lemma A.1 relies on multiple iterations of the chain rule for mutual information:

$$\begin{aligned} I(Z,S) &\overset{(a)}{=} I(Z,\{X,S\}) - I(Z,X|S) \\ &\overset{(b)}{=} I(Z,X) + I(Z,S|X) - I(Z,X|S) \\ &\overset{(c)}{=} I(Z,X) - I(Z,X|S) \\ &\overset{(d)}{=} I(Z,X) - I(X,\{Z,S\}) + I(X,S) \\ &\overset{(e)}{=} I(Z,X) + H(X|Z,S) - H(X|S) \end{aligned}$$

where (a), (b) and (d) use the chain rule for mutual information; and, (c) uses the fact that Z is only encoded from X and from S, so $H(Z|X,S) = H(Z|X)$ and $I(Z,S|X) = H(Z|X) - H(Z|X,S) = 0$. And (e) uses the fact that $I(X,S) = H(X) - H(X|S)$ and $I(X,\{Z,S\}) = H(X) - H(X|Z,S)$. □

Lemma A.1 implies that if the distortion is $d(X,\{Z,S\}) = H(X|Z,S)$, the unfairness-distortion function is given by

$$\begin{aligned} I(D) = &\min_f I(Z,X) + H(X|Z,S) - H(X|S) \\ &\text{s.t. } H(X,\{Z,S\}) \leq D \end{aligned} \tag{6}$$

Second, a fundamental theorem in rate-distortion [Cover and Thomas, 2012] shows that if the distortion is $d(X,\{Z,S\}) = H(X|Z,S)$ the rate-distortion function is given by

$$R(D) = \min_f I(X,Z) \text{ s.t } H(X|Z,S) \leq D, \tag{7}$$

and that $R(D)$ is a non-increasing convex function. The next Lemma shows how solution of the minimization problem (7) solves the minimization problem (6) whenever $\frac{\partial R(D)}{\partial D} \leq -1$

Lemma A.2. *Let $D \geq 0$ be a distortion value. Assume that $\frac{\partial R(D)}{\partial D} \leq -1$. A solution f^* of the minimization* (7) *for D is also solution of* (6)*.*

Proof. At the optimum, the constraint in (7) is binding and thus, that $H_{f^*}(X|Z,S) = D$, where the subs-script f^* reminds that the code Z depends on f^*. Consider now a solution g^* of the minimization (6) for a distortion D. We consider two cases: case (I) the constraint is binding for g^* in (6); case (II) the constraint is not binding for g^* in (6).

case (I): $H_{g^*}(X|Z,S) = D$ and we have

$$\begin{aligned} I(D) &= I_{g^*}(Z,X) + H_{g^*}(X|Z,S) - H(X|S) \\ &= I_{g^*}(Z,X) + D - H(X|S) \\ &\overset{(a)}{\geq} I_{f^*}(Z,X) + D - H(X|S), \end{aligned} \tag{8}$$

where (a) uses the fact that f^* is solution of (7) and that $H_{g^*}(X|Z,S) \leq D$. Therefore, since $H_{f^*}(X|Z,S) \leq D$, f^* is also solution of (6).

case (II): Let denote $D^{'}$ the value of the distortion achieved by g^*. Then, $D^{'} = H_{g^*}(X|Z,S) < D$. We have

$$\begin{aligned} I(D) &= I_{g^*}(Z,X) + H_{g^*}(X|Z,S) - H(X|S) \\ &= I_{g^*}(Z,X) + D^{'} - H(X|S) \\ &\overset{(a)}{\geq} R(D^{'}) + D^{'} - H(X|S), \end{aligned} \tag{9}$$

where (a) follows from the definition of $R(D^{'})$. By convexity of the rate-distortion function, we have that

$$\begin{aligned} R(D^{'}) - R(D) &\overset{(a)}{\geq} \frac{\partial R(D)}{\partial D}(D^{'} - D) \\ &\overset{(b)}{\geq} (D - D^{'}), \end{aligned} \tag{10}$$

where (a) uses the convexity of $R(D)$ and that $D^{'} < D$ and (b) uses that $\frac{\partial R(D)}{\partial D} \leq -1$. Hence, by combining (9) and (10), we have

$$I(D) \geq R(D) + D - H(X|S) = I_{f^*}(Z,X) + D - H(X|S).$$

Therefore, f^* is also solution of the minimization (6) since $H_{f^*}(X|Z,S) \leq D$. □

It follows from Lemma A.2 that we have by definition of f^*, if $\frac{\partial R(D)}{\partial D} \leq -1$

$$I(D) = I_{f^*}(Z,X) + D - H(X|S) = R(D) + D - H(X|S),$$

which proves the first part of the statement in Theorem 2.1. Moreoover, if $\frac{\partial R(D)}{\partial D} < -1$, $\frac{\partial I(D)}{\partial D} = \frac{\partial R(D)}{\partial D} + 1 < 0$, hence $I(.)$ is decreasing for D such that $\frac{\partial R(D)}{\partial D} < -1$.

To prove that if $\frac{\partial R(D)}{\partial D} \geq -1$, $I(D) = 0$, we first prove the following Lemma:

Lemma A.3. *Let D^* denote the value of D such that $\frac{\partial R(D)}{\partial D} = -1$. For $D^* \geq D$, $I(D) = I(D^*)$.*

Proof. Let $D > D^*$. Let g^* be a solution of the minimization (6) for D. Note that a solution of (6) for D^* respects the constraint of the minimization (6) for D and thus, $I(D^*) \geq I(D)$. Let $D^{'}$ denote $H_{g^*}(X|Z,S)$. Then, by definition of the rate-distortion objective value (7), we have

$$\begin{aligned} I(D) &= I_{g^*}(Z,X) + D^{'} - H(X|S) \\ &\geq R(D^{'}) + D^{'} - H(X|S). \end{aligned} \tag{11}$$

If $D^{'} < D^*$, then we already know that $I(D^{'}) = R(D^{'}) + D^{'} - H(X|S)$ and that $I(D^{'}) > I(D^*) \geq I(D)$. Moreover, by inequality (11), $\geq I(D^{'})$, thus $I(D^{'}) > I(D) \geq I(D^{'})$, which is a contradiction. If $D^{'} = D^*$, we already know that $I(D) \leq I(D^*) = R(D^*) + D^* - H(X|S) = I(D^{'}) \leq I(D)$ and thus that $I(D) = I(D^*)$.

It remains to look at the case $D^{'} > D^*$. Consider $D^{"} \in [D^*, D^{'}]$. By convexity of $R(D)$ we have

$$R(D^*) - R(D^{'}) \leq \frac{\partial R(D^*)}{\partial D}(D^* - D^{'}) \overset{(a)}{=} D^{'} - D^*,$$

where (a) comes the fact that $\frac{\partial R(D^*)}{\partial D} = -1$. It results that by the inequality (11) $I(D) \geq R(D^*) + D^* - H(X|S)$. Moreover, we already know that $R(D^*) + D^* - H(X|S) = I(D^*)$. Hence $I(D^*) \geq I(D) \geq I(D^*)$, which proves the equality in Lemma A.2. □

Lemma A.4. *Let $D^{**} = H(X|S)$. We have $I(D^{**}) = 0$.*

Proof. Consider an encoder g that generates a random variable Z independent of X. Then $H_g(X|Z,S) = D^{**}$ and $I_g(Z,X) = 0$. Therefore, g respect the constraint of the minimization (6) for D^{**} and $I(D^{**}) \leq I_g(Z,X) + H_g(X|Z,S) - H(X|S) = 0$. Hence, $I(D^{**}) = 0$. □

By combining Lemma A.2 and A.4, we can show that $I(D) = 0$ for $D \geq D^{**}$.

A.2 Entropy Estimation

We follow a standard approach in rate-distortion [Theis *et al.*, 2017; Choi *et al.*, 2019] and approximate the discrete distribution $p(z_j|z_{.<j}, \beta)$ by a continuous distribution $q(z_j|z_{.<j}, \beta)$ such that the probability mass of q on the interval $[z_j - 1/2^{a_j(\beta)}, z_j + 1/2^{a_j(\beta)}]$ is equal to $p(z_j|z_{.<j}, \beta)$. Therefore,

$$\begin{aligned} H(z|\beta) &= -\sum_{j=1}^{d} E\left[\log p(z_j|z_{.<j}, \beta)\right] \\ &= -\sum_{j=1}^{d} E\left[\log\left(\int_{\frac{-1}{2^{a_j(\beta)}}}^{\frac{1}{2^{a_j(\beta)}}} q(z_j + u|z_{.<j}, \beta)du\right)\right] \\ &+ KL\left(p||\int_{\frac{-1}{2^{a_j(\beta)}}}^{\frac{1}{2^{a_j(\beta)}}} q(z_j + u|z_{.<j}, \beta)du\right) \\ &\overset{(a)}{\leq} -\sum_{j=1}^{d} E\left[\log\left(\int_{\frac{-1}{2^{a_j(\beta)}}}^{\frac{1}{2^{a_j(\beta)}}} q(z_j + u|z_{.<j}, \beta)du\right)\right] \end{aligned} \tag{12}$$

where (a) uses the non-negativity of the Kullback-Leibler divergence KL between the true distribution $p(z|\beta)$ and its approximation $q(z|\beta)$ once convolved with a uniform distribution over $[-1/2^{a_j(\beta)}, 1/2^{a_j(\beta)}]$.

We follow [Salimans *et al.*, 2017] and for each $j = 1, ..., d$ we model $q(.|z_{.<j}, \beta)$ as a mixture of K logistic distributions with means $\mu_{j,k}(\beta)$, scales $\gamma_{j,k}(\beta)$ and mixtures probability $\pi_{j,k}(\beta)$, which allows to compute exactly the integral term in (12). Specifically, we compute

$$\mu_{j,k} = \mu^0_{j,k}(\beta) + w^{\mu}_{j,k}(\beta)\Gamma_j \odot z_j, \tag{13}$$

and

$$\log(\gamma_{j,k}) = \gamma^0_{j,k}(\beta) + w^{\gamma}_{j,k}(\beta)\Gamma_j \odot z_j, \tag{14}$$

where $\mu^0_{j,k}(.), \gamma^0_{j,k}()$ are functions from $[0,1]$ to $\mathbb{R}$; $w^{\mu}_{jk}()$ and $w^{\gamma}_{j,k}()$ are functions from $[0,1]$ to $\mathbb{R}^d$; and, $\Gamma_j = (1, 1, .., 1, 0, ...0)$ is a $d-$ dimensional vector equal to one for entry before j and zero otherwise. Γ_j guarantees that the distribution $q(.|z_{.<j})$ is conditioned only on $z_{.<j}$and not on any $z_{j^{'}}$ for $j^{'} \geq j$.

The use of logistic distribution allows to compute the upper bound in (12) as $H_q(z|\beta)$ where $H_q(z|\beta)$ is given by

$$-\sum_{j=1}^{d} E\left[\log\left(\sum_{k=1}^{K} \pi_{j,k}\sigma\left(\frac{z_j + \mu_{j,k}(\beta)}{\gamma_{j,k}(\beta)} + \frac{1}{2^{a_j(\beta)}}\right) - \sigma\left(\frac{z_j + \mu_{j,k}(\beta)}{\gamma_{j,k}(\beta)} - \frac{1}{2^{a_j(\beta)}}\right)\right)\right].$$

Acknowledgements

This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

References

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, 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, 23–24 Feb 2018. PMLR.

[Choi *et al.*, 2019] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee. Variable rate deep image compression with a conditional autoencoder. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3146–3154, 2019.

[Chouldechova and Roth, 2018] Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

[Cover and Thomas, 2012] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[Creager *et al.*, 2019] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

[Cui *et al.*, 2020] Ze Cui, Jing Wang, Bo Bai, Tiansheng Guo, and Yihui Feng. G-vae: A continuously variable rate deep image compression framework. *arXiv preprint arXiv:2003.02012*, 2020.

[Edwards and Storkey, 2016] Harrison Edwards and Amos J. Storkey. Censoring representations with an adversary. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[Gardner *et al.*, 2019] Josh Gardner, Christopher Brooks, and Ryan Baker. 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, 2019.

[Gitiaux and Rangwala, 2021a] Xavier Gitiaux and Huzefa Rangwala. Fair representations by compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11506–11515, 2021.

[Gitiaux and Rangwala, 2021b] Xavier Gitiaux and Huzefa Rangwala. Learning smooth and fair representations. In *International Conference on Artificial Intelligence and Statistics*, pages 253–261. PMLR, 2021.

[Gupta *et al.*, 2021] Umang Gupta, Aaron Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7610–7619, 2021.

[Higgins *et al.*, 2016] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[Jaiswal *et al.*, 2020] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4272–4279, 2020.

[Kostina and Tuncel, 2019] Victoria Kostina and Ertem Tuncel. Successive refinement of abstract sources. *IEEE Transactions on Information Theory*, 65(10):6385–6398, 2019.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[Moyer *et al.*, 2018] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pages 9084–9093, 2018.

[Pfohl *et al.*, 2019] Stephen Pfohl, Ben Marafino, Adrien Coulet, Fatima Rodriguez, Latha Palaniappan, and Nigam H Shah. 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, 2019.

[ProPublica, 2016] ProPublica. How we analyzed the compas recidivism algorithm. *ProPublica*, 2016.

[Roy and Boddeti, 2019] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2594, 2019.

[Salimans *et al.*, 2017] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv preprint arXiv:1701.05517*, 2017.

[Theis *et al.*, 2017] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*, 2017.

[Tishby, 1999] N Tishby. The information bottleneck method. In *Proc. 37th Annual Allerton Conference on Communications, Control and Computing, 1999*, pages 368–377, 1999.

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