Flexibly Fair Representation Learning by Disentanglement

Elliot Creager [1 2] **David Madras** [1 2] **Jörn-Henrik Jacobsen** [2] **Marissa A. Weis** [2 3] **Kevin Swersky** [4]
Toniann Pitassi [1 2] **Richard Zemel** [1 2]

Abstract

We consider the problem of learning representations that achieve group and subgroup fairness with respect to multiple sensitive attributes. Taking inspiration from the disentangled representation learning literature, we propose an algorithm for learning compact representations of datasets that are useful for reconstruction and prediction, but are also *flexibly fair*, meaning they can be easily modified at test time to achieve subgroup demographic parity with respect to multiple sensitive attributes and their conjunctions. We show empirically that the resulting encoder—which does not require the sensitive attributes for inference—enables the adaptation of a single representation to a variety of fair classification tasks with new target labels and subgroup definitions.

1. Introduction

Machine learning systems are capable of exhibiting discriminatory behaviors against certain demographic groups in high-stakes domains such as law, finance, and medicine (Kirchner et al., 2016; Aleo & Svirsky, 2008; Kim et al., 2015). These outcomes are potentially unethical or illegal (Barocas & Selbst, 2014; Hellman, 2018), and behoove researchers to investigate more equitable and robust models. One promising approach is fair representation learning: the design of neural networks using learning objectives that satisfy certain fairness or parity constraints in their outputs (Zemel et al., 2013; Louizos et al., 2015; Edwards & Storkey, 2015; Madras et al., 2018). This is attractive because neural network representations often generalize to tasks that are unspecified at training time, which implies that a properly specified fair network can act as a group parity bottleneck that reduces discrimination in unknown downstream tasks.

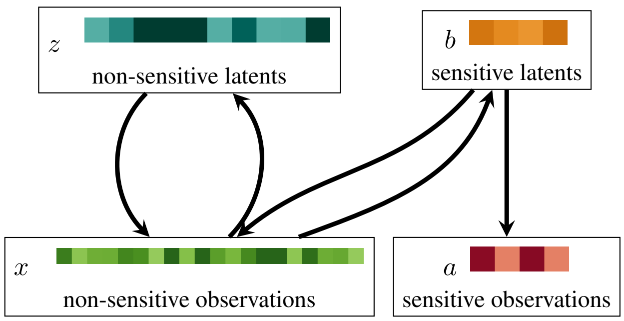
[1]University of Toronto [2]Vector Institute [3]University of Tübingen [4]Google Research. Correspondence to: Elliot Creager <creager@cs.toronto.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

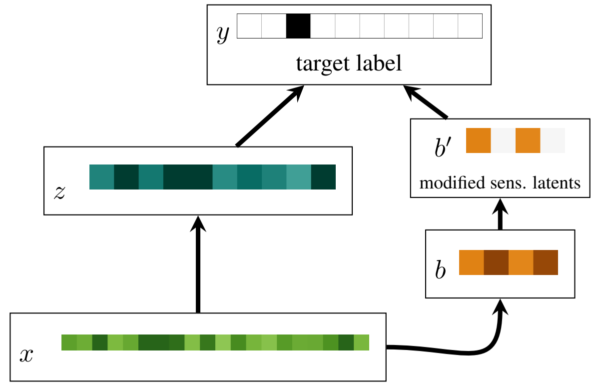
Current approaches to fair representation learning are flexible with respect to downstream tasks but inflexible with respect to sensitive attributes. While a single learned representation can adapt to the prediction of different task labels y, the single sensitive attribute a for *all* tasks must be specified at train time. Mis-specified or overly constraining train-time sensitive attributes could negatively affect performance on downstream prediction tasks. Can we instead learn a *flexibly fair* representation which can be adapted, at test time, to be fair to a variety of protected groups and their intersections? Such an adaptation procedure should satisfy two criteria. Firstly, the latent space should be structured so that the adaption is *simple*, allowing a practitioner to easily adapt the representation to a variety of fair classification settings, where each task may have a different task label y and sensitive attributes a. Additionally, we require the adaptations are *compositional*: the representations can be made fair with respect to conjunctions of sensitive attributes, to guard against subgroup discrimination, for example, a classifier which is fair to women, but not Black women over the age of 60. This type of subgroup discrimination has been observed in commercial machine learning systems (Buolamwini & Gebru, 2018).

In this work, we investigate how to learn flexibly fair representations that can be easily adapted at test time to achieve fairness with respect to sets of sensitive groups or subgroups. We draw inspiration from the disentangled representation literature, where the goal is for each dimension of the representation (also called the "latent code") to correspond to no more than one semantic factor of variation in the data (for example, independent visual features like object shape and position) (Higgins et al., 2016; Locatello et al., 2018). In our method, we use multiple sensitive attribute labels at train time to induce a disentangled structure in the learned representation, which allows us to easily eliminate their influence at test time. Importantly, at test time our method does not require access to the sensitive attributes, which can be difficult to collect in practice due to legal restrictions (Elliot et al., 2008; DCCA, 1983). The trained representation permits simple and composable modifications at test time that eliminate the influence of sensitive attributes, enabling a wide variety of downstream tasks.

We first provide a proof-of-concept by generating a variant



(a) FFVAE learns the encoder distribution $q(z, b|x)$ and decoder distributions $p(x|z, b)$, $p(a|b)$ from inputs x and multiple sensitive attributes a. The disentanglement prior structures the latent space by encouraging low $\mathrm{MI}(b_i, a_j) \forall i \neq j$ and low $\mathrm{MI}(b, z)$ where $\mathrm{MI}(\cdot)$ denotes mutual information.



(b) The FFVAE latent code $[z, b]$ can be modified by discarding or noising out sensitive dimensions $\{b_j\}$, which yields a latent code $[z, b']$ independent of groups and subgroups derived from sensitive attributes $\{a_j\}$. A held out label y can then be predicted with subgroup demographic parity.

Figure 1. Data flow at train time (1a) and test time (1b) for our model, Flexibly Fair VAE (FFVAE).

of the synthetic DSprites dataset with correlated ground truth factors of variation, which is better suited to fairness-related questions. We demonstrate that even in the correlated setting, our method is capable of disentangling the effect of several sensitive attributes from data, and that this disentanglement is useful for fair classification tasks downstream. We then apply our method to a real-world tabular dataset (Communities & Crime) where we find that our method matches or exceeds the fairness-accuracy tradeoff of existing disentangled representation learning approaches on a majority of the evaluated subgroups.

2. Background

Group Fair Classification In fair classification, we consider labeled examples $x, a, y \sim p_{\text{data}}$ where $y \in \mathcal{Y}$ are labels we wish to predict, $a \in \mathcal{A}$ are *sensitive* attributes, and $x \in \mathcal{X}$ are non-sensitive attributes. The goal is to learn a classifier $\hat{y} = g(x, a)$ (or $\hat{y} = g(x)$) which is predictive of y and achieves certain group fairness criteria w.r.t. a. These criteria are typically written as independence properties of the various random variables involved. In this paper we focus on demographic parity, which is satisfied when the predictions are independent of the sensitive attributes: $\hat{y} \perp a$. It is often impossible or undesirable to satisfy demographic parity exactly (i.e. achieve complete independence). In this case, a useful metric is *demographic parity distance*:

$$\Delta_{DP} = |\mathbb{E}[\bar{y} = 1|a = 1] - \mathbb{E}[\bar{y} = 1|a = 0]| \qquad (1)$$

where $\bar{y}$ here is a binary prediction derived from model output $\hat{y}$. When $\Delta_{DP} = 0$, demographic parity is achieved; in general, lower Δ_{DP} is associated with more fair classification.

Our work differs from the fair classification setup in two ways. Firstly, we consider several sensitive attributes at once, and aim to enable fair classification with respect to each; individually as well as jointly (cf. subgroup fair classification, Kearns et al. (2017); Hebert-Johnson et al. (2018)). Secondly, we focus on representation learning rather than classification, with the aim of enabling a range of fair classification tasks downstream.

Group Fair Representation Learning In order to flexibly deal with many different label and sensitive attribute sets, we employ representation learning to compute a compact but predicatively useful encoding of the dataset that can be flexibly adapted to different fair classification tasks. As an example, if we learn the function f that achieves independence in the representations $z \perp a$ with $z = f(x, a)$ or $z = f(x)$, then any predictor derived from this representation will also achieve the desired demographic parity, $\hat{y} \perp a$ with $\hat{y} = g(z)$.

The fairness literature typically considers binary labels and sensitive attributes: $\mathcal{A} = \mathcal{Y} = \{0, 1\}$. In this case, approaches like regularization (Zemel et al., 2013) and adversarial regularization (Edwards & Storkey, 2015; Madras et al., 2018) are straightforward to implement. We want to address the case where a is a vector with many dimensions. Group fairness must be preserved with respect to each dimension of a (age, race, gender, etc.) and their combinations.

VAE The vanilla Variational Autoencoder (VAE) (Kingma & Welling, 2013) is typically implemented with an isotropic Gaussian prior $p(z) = \mathcal{N}(0, I)$. The objective to be maximized is the Evidence Lower Bound (a.k.a., the ELBO),

$$L_{\text{VAE}}(p, q) = \mathbb{E}_{q(z|x)} [\log p(x|z)] - D_{KL} [q(z|x)||p(z)], \tag{2}$$

which bounds the data log likelihood $\log p(x)$ from below for any choice of q. The encoder and decoder are often implemented as Gaussians

$$\begin{aligned} q(z|x) &= \mathcal{N}(z|\mu_q(x), \Sigma_q(x)) \\ p(x|z) &= \mathcal{N}(x|\mu_p(z), \Sigma_p(z)) \end{aligned} \tag{3}$$

whose distributional parameters are the outputs of neural networks $\mu_q(\cdot)$, $\Sigma_q(\cdot)$, $\mu_p(\cdot)$, $\Sigma_p(\cdot)$, with the Σ typically exhibiting diagonal structure. For modeling binary-valued pixels, a Bernoulli decoder $p(x|z) = \text{Bernolli}(x|\theta_p(z))$ can be used. The goal is to maximize L_{VAE}—which is made differentiable by reparameterizing samples from $q(z|x)$—w.r.t. the network parameters.

β-VAE Higgins et al. (2016) modify the VAE objective:

$$L_{\beta\text{VAE}}(p, q) = \mathbb{E}_{q(z|x)} [\log p(x|z)] - \beta D_{KL} [q(z|x)||p(z)]. \tag{4}$$

The hyperparameter β allows the practitioner to encourage the variational distribution $q(z|x)$ to reduce its KL-divergence to the isotropic Gaussian prior $p(z)$. With $\beta > 1$ this objective is a valid lower bound on the data likelihood. This gives greater control over the model's adherence to the prior. Because the prior factorizes per dimension $p(z) = \prod_j p(z_j)$, Higgins et al. (2016) argue that increasing β yields "disentangled" latent codes in the encoder distribution $q(z|x)$. Broadly speaking, each dimension of a properly disentangled latent code should capture no more than one semantically meaningful factor of variation in the data. This allows the factors to be manipulated in isolation by altering the per-dimension values of the latent code. Disentangled autoencoders are often evaluated by their sample quality in the data domain, but we instead emphasize the role of the encoder as a representation learner to be evaluated on downstream fair classification tasks.

FactorVAE and β-TCVAE Kim & Mnih (2018) propose a different variant of the VAE objective:

$$L_{\text{FactorVAE}}(p, q) = L_{\text{VAE}}(p, q) - \gamma D_{KL}(q(z)||\prod_j q(z_j)). \tag{5}$$

The main idea is to encourage factorization of the aggregate posterior $q(z) = \mathbb{E}_{p^{\text{data}}(x)} [q(z|x)]$ so that z_i correlates with z_j if and only if $i = j$. The authors propose a simple trick to generate samples from the aggregate posterior $q(z)$ and its marginals $\{q(z_j)\}$ using shuffled minibatch indices, then approximate the $D_{KL}(q(z)||\prod_j q(z_j))$ term using the cross entropy loss of a classifier that distinguishes between the two sets of samples, which yields a mini-max optimization.

Chen et al. (2018) show that the $D_{KL}(q(z)||\prod_j q(z_j))$ term in Equation 5—a.k.a. the "total correlation" of the latent code—can be naturally derived by decomposing the expected KL divergence from the variational posterior to prior:

$$\begin{aligned} \mathbb{E}_{p^{\text{data}}(x)}[D_{KL}(q(z|x)||p(z))] = \\ D_{KL}(q(z|x)p^{\text{data}}(x)||q(z)p^{\text{data}}(x)) \\ + D_{KL}(q(z)||\prod_j q(z_j)) \\ + \sum_j D_{KL} [q(z_j)||p(z_j)]. \end{aligned} \tag{6}$$

They then augment the decomposed ELBO to arrive at the same objective as Kim & Mnih (2018), but optimize using a biased estimate of the marginal probabilities $q(z_j)$ rather than with the adversarial bound on the KL between aggregate posterior and its marginals.

3. Related Work

Most work in fair machine learning deals with fairness with respect to single (binary) sensitive attributes. Multi-attribute fair classification was recently the focus of Kearns et al. (2017)—with empirical follow-up (Kearns et al., 2018)—and Hebert-Johnson et al. (2018). Both papers define the notion of an identifiable class of subgroups, and then obtain fair classification algorithms that are provably as efficient as the underlying learning problem for this class of subgroups. The main difference is the underlying metric; Kearns et al. (2017) use statistical parity whereas Hebert-Johnson et al. (2018) focus on calibration. Building on the multi-accuracy framework of Hebert-Johnson et al. (2018), Kim et al. (2018) develop a new algorithm to achieve multi-group accuracy via a post-processing boosting procedure.

The search of independent latent components that explain observed data has long been a focus on the probabilistic modeling community (Comon, 1994; Hyvärinen & Oja, 2000; Bach & Jordan, 2002). In light of the increased prevalence of neural networks models in many data domains, the machine learning community has renewed its interest in learned features that "disentangle" semantic factors of data variation. The introduction of the β-VAE (Higgins et al., 2016), as discussed in section 2, motivated a number of subsequent studies that examine why adding additional weight on the KL-divergence of the ELBO encourages disentangled representations (Alemi et al., 2018; Burgess et al., 2018). Chen et al. (2018); Kim & Mnih (2018) and Esmaeili et al. (2018) argue that decomposing the ELBO and penalizing the total correlation increases disentanglement in the latent representations. Locatello et al. (2018) conduct extensive experiments comparing existing unsupervised disentanglement methods and metrics. They conclude pessimistically that learning disentangled representations requires inductive biases and possibly additional supervision, but identify fair machine learning as a potential application where additional supervision is available by way of sensitive attributes.

Our work is the first to consider multi-attribute fair representation learning, which we accomplish by using sensitive attributes as labels to induce a factorized structure in the aggregate latent code. Bose & Hamilton (2018) proposed a compositional fair representation of graph-structured data. Kingma et al. (2014) previously incorporated (partially-observed) label information into the VAE framework to perform semi-supervised classification. Several recent VAE variants have incorporated label information into latent variable learning for image synthesis (Klys et al., 2018) and single-attribute fair representation learning (Song et al., 2018; Botros & Tomczak, 2018; Moyer et al., 2018). Designing invariant representations with non-variational objectives has also been explored, including in reversible models (Ardizzone et al., 2018; Jacobsen et al., 2018).

4. Flexibly Fair VAE

We want to learn fair representations that—beyond being useful for predicting many test-time task labels y—can be adapted *simply* and *compositionally* for a variety of sensitive attributes settings a after training. We call this property *flexible fairness*. Our approach to this problem involves inducing structure in the latent code that allows for easy manipulation. Specifically, we isolate information about each sensitive attribute to a specific subspace, while ensuring that the latent space factorizes these subspaces independently.

Notation We employ the following notation:

- $x \in \mathcal{X}$: a vector of non-sensitive attributes, for example, the pixel values in an image or row of features in a tabular dataset;
- $a \in \{0,1\}^{N_a}$: a vector of binary sensitive attributes;
- $z \in \mathbb{R}^{N_z}$: the non-sensitive dimensions of the latent code;
- $b \in \mathbb{R}^{N_b}$: the sensitive dimensions of the latent code[1].

[1] In our experiments we used $N_b = N_a$—the same number of sensitive attributes as sensitive latent code dimensions—to model binary sensitive attributes. But categorical or continuous sensitive attributes can also be accommodated.

For example, we can express the VAE objective in this notation as

$$L_{\text{VAE}}(p,q) = \mathbb{E}_{q(z,b|x,a)}\left[\log p(x,a|z,b)\right] - D_{KL}\left[q(z,b|x,a)||p(z,b)\right]. \qquad (7)$$

In learning a flexibly fair representations $[z,b] = f([x,a])$, we aim to satisfy two general properties: *disentanglement* and *predictiveness*. We say that $[z,b]$ is *disentangled* if its aggregate posterior factorizes as $q(z,b) = q(z)\prod_j q(b_j)$ and is *predictive* if each b_i has high mutual information with the corresponding a_i. Note that under the disentanglement criteria the dimensions of z are free to co-vary together, but must be independent from all sensitive subspaces b_j. We have also specified factorization of the latent space in terms of the aggregate posterior $q(z,b) = \mathbb{E}_{p^{\text{data}}(x)}[q(z,b|x)]$, to match the global independence criteria of group fairness.

Desiderata We can formally express our desiderata as follows:

- $z \perp b_j \; \forall\, j$ (disentanglement of the non-sensitive and sensitive latent dimensions);
- $b_i \perp b_j \; \forall\, i \neq j$ (disentanglement of the various different sensitive dimensions);
- $\text{MI}(a_j, b_j)$ is large $\forall\, j$ (predictiveness of each sensitive dimension);

where $\text{MI}(u,v) = \mathbb{E}_{p(u,v)} \log \frac{p(u,v)}{p(u)p(v)}$ represents the mutual information between random vectors u and v. We note that these desiderata differ in two ways from the standard disentanglement criteria. The predictiveness requirements are stronger: they allow for the injection of external information into the latent representation, requiring the model to structure its latent code to align with that external information. However, the disentanglement requirement is less restrictive since it allows for correlations between the dimensions of z. Since those are the non-sensitive dimensions, we are not interested in manipulating those at test time, and so we have no need for constraining them.

If we satisfy these criteria, then it is possible to achieve demographic parity with respect to some a_i by simply removing the dimension b_i from the learned representation i.e. use instead $[z,b]\backslash b_i$. We can alternatively replace b_i with independent noise. This adaptation procedure is simple and compositional: if we wish to achieve fairness with respect to a conjunction of binary attributes[2] $a_i \wedge a_j \wedge a_k$, we can simply use the representation $[z,b]\backslash\{b_i, b_j, b_k\}$.

By comparison, while an objective such as FactorVAE may disentangle dimensions of the aggregate posterior—$q(z) = \prod_j q(z_j)$—it does not automatically satisfy flexible

[2] $\wedge$ and $\vee$ represent logical *and* and *or* operations, respectively.

fairness, since the representations are not predictive, and cannot necessarily be easily modified along the attributes of interest.

Distributions We propose a variation to the VAE which encourages our desiderata, building on methods for disentanglement and encouraging predictiveness. Firstly, we assume assume a variational posterior that factorizes across z and b:

$$q(z, b|x) = q(z|x)q(b|x). \tag{8}$$

The parameters of these distributions are implemented as neural network outputs, with the encoder network yielding a tuple of parameters for each input: $(\mu_q(x), \Sigma_q(x), \theta_q(x)) =$ Encoder(x). We then specify $q(z|x) = \mathcal{N}(z|\mu_q(x), \Sigma_q(x))$ and $q(b|x) = \delta(\theta_q(x))$ (i.e., b is non-stochastic)[3].

Secondly, we model reconstruction of x and prediction of a separately using a factorized decoder:

$$p(x, a|z, b) = p(x|z, b)p(a|b) \tag{9}$$

where $p(x|z, b)$ is the decoder distribution suitably chosen for the inputs x, and $p(a|b) = \prod_j \text{Bernoulli}(a_j|\sigma(b_j))$ is a factorized binary classifier that uses b_j as the logit for predicting a_j ($\sigma(\cdot)$ represents the sigmoid function here). Note that the $p(a|b)$ factor of the decoder requires no extra parameters.

Finally, we specify a factorized prior $p(z, b) = p(z)p(b)$ with $p(z)$ as a standard Gaussian and $p(b)$ as Uniform.

Learning Objective Using the encoder and decoder as defined above, we present our final objective:

$$\begin{aligned} L_{\text{FFVAE}}(p, q) = \mathbb{E}_{q(z,b|x)}[\log p(x|z, b) + \alpha \log p(a|b)] \\ - \gamma D_{KL}(q(z, b)||q(z) \prod_j q(b_j)) \\ - D_{KL}\left[q(z, b|x)||p(z, b)\right]. \end{aligned} \tag{10}$$

It comprises the following four terms, respectively: a *reconstruction* term which rewards the model for successfully modeling non-sensitive observations; a *predictiveness* term which rewards the model for aligning the correct latent components with the sensitive attributes; a *disentanglement* term which rewards the model for decorrelating the latent dimensions of b from each other and z; and a *dimension-wise KL* term which rewards the model for matching the prior in the latent variables. We call our model FFVAE for Flexibly Fair VAE (see Figure 1 for a schematic representation).

The hyperparameters α and γ control aspects relevant to flexible fairness of the representation. α the alignment of each a_j to its corresponding b_j (predictiveness), whereas γ controls the aggregate independence in the latent code (disentanglement).

The γ-weighted total correlation term is realized by training a binary adversary to approximate the log density ratio $\log \frac{q(z,b)}{q(z)\prod_j q(b_j)}$. The adversary attempts to classify between "true" samples from the aggregate posterior $q(z, b)$ and "fake" samples from the product of the marginals $q(z) \prod_j q(b_j)$ (see Appendix A for further details). If a strong adversary can do no better than random chance, then the desired independence property has been achieved.

We note that our model requires the sensitive attributes a at training time but not at test time. This is advantageous, since often these attributes can be difficult to collect from users, due to practical and legal restrictions, particularly for sensitive information (Elliot et al., 2008; DCCA, 1983).

5. Experiments

5.1. Evaluation Criteria

We evaluate the learned encoders with an "auditing" scheme on held-out data. The overall procedure is as follows:

1. **Split data** into a *training* set (for learning the encoder) and an *audit* set (for evaluating the encoder).

2. **Train** an encoder/representation using the training set.

3. **Audit** the learned encoder. Freeze the encoder weights and train an MLP to predict some task label given the (possibly modified) encoder outputs on the audit set.

To evaluate various properties of the encoder we conduct three types of auditing tasks—fair classification, predictiveness, and disentanglement—which vary in task label and representation modification. The fair classification audit (Madras et al., 2018) trains an MLP to predict y (held-out from encoder training) given $[z, b]$ with appropriate sensitive dimensions removed, and evaluates accuracy and Δ_{DP} on a test set. We repeat for a variety of demographic groups and subgroups derived from the sensitive attributes. The predictiveness audit trains classifier C_i to predict sensitive attribute a_i from b_i alone. The disentanglement audit trains classifier $C_{\setminus i}$ to predict sensitive attribute a_i from the representation with b_i removed (e.g. $[z, b]\setminus b_i$). If C_i has low loss, our representation is predictive; if $C_{\setminus i}$ has high loss, it is disentangled.

[3] We experimented with several distributions for modeling $b|x$ stochastically, but modeling this uncertainty did not help optimization or downstream evaluation in our experiments.

5.2. Synthetic Data

DSpritesUnfair Dataset The DSprites dataset[4] contains 64×64-pixel images of white shapes against a black background, and was designed to evaluate whether learned representations have disentangled sources of variation. The original dataset has several categorical factors of variation—Scale, Orientation, XPosition, YPosition—that combine to create $700,000$ unique images. We binarize the factors of variation to derive sensitive attributes and labels, so that many images now share any given attribute/label combination (See Appendix B for details). In the original DSprites dataset, the factors of variation are sampled uniformly. However, in fairness problems, we are often concerned with correlations between attributes and the labels we are trying to predict (otherwise, achieving low Δ_{DP} is aligned with standard classification objectives). Hence, we sampled an "unfair" version of this data (DSpritesUnfair) with correlated factors of variation; in particular Shape and XPosition correlate positively. Then a non-trivial fair classification task would be, for instance, learning to predict shape without discriminating against inputs from the left side of the image.

Baselines We compare our model against several baselines. To test the utility of our predictiveness prior, two of these baselines are the β-VAE (VAE with a coefficient $\beta \geq 1$ on the KL term) and FactorVAE, which have disentanglement priors but no predictiveness prior. We can also think of these as FFVAE with $\alpha = 0$. To test the utility of our disentanglement prior, we also compare against a version of our model with $\gamma = 0$. This is similar to the class-conditional VAE (Kingma et al., 2014), with sensitive attributes as labels — this model encourages predictiveness but no disentanglement.

Fair Classification We perform the fair classification audit using several group/subgroup definitions for models trained on DSpritesUnfair (see Appendix D for details), and report fairness-accuracy tradeoff curves in Fig. 2. In these experiments, we used Shape and Size as our sensitive attributes during encoder training. We perform the fair classification audit, predicting the label $y =$"XPosition"—which was not used in the representation learning phase—given the modified encoder outputs, and repeat for several sensitive groups and subgroups. We modify the encoder outputs as follows: When our sensitive attribute is a_i we remove the associated dimension b_i from $[z, b]$; when the attribute is a conjunction of a_i and a_j, we remove both of b_i and b_j. For the β-VAE baseline, we simply remove the latent dimension which is most correlated with a_i, or the two most correlated dimensions with the conjunction. We sweep a range of hyperparameters to produce the fairness-accuracy tradeoff curve for each model. In Fig. 2, we show the "Pareto front" of these models: points in (Δ_{DP}, accuracy)-space for which no other point is better along both dimensions. The optimal result is the top left hand corner (perfect accuracy and $\Delta_{DP} = 0$).

Since we have a 2-D sensitive input space, we show results for four different sensitive attributes derived from these dimensions: $\{a =$ "Shape"$, a =$ "Size"$, a =$ "Shape" $\vee$ "Size"$, a =$ "Shape" $\wedge$ "Size"$\}$. Recall that Shape and XPosition correlate in the DSpritesUnfair dataset. Therefore, for sensitive attributes that involve Shape, we expect to see an improvement in Δ_{DP}. For sensitive attributes that do not involve Shape, we expect that our method does not hurt performance at all — since the attributes are uncorrelated in the data, the optimal predictive solution also has $\Delta_{DP} = 0$.

In Fig. 2, we observe this behaviour. In Fig. 2a, our sensitive attribute Scale is completely uncorrelated with y. Therefore, we expect all our models to achieve high accuracy and low Δ_{DP} on this problem, which they do. This experiment functions as a sanity check — when we remove the dimension for Shape from our learned representation, it doesn't hurt our audit's accuracy at all. This means what we have completely disentangled the information in Shape from the label information.

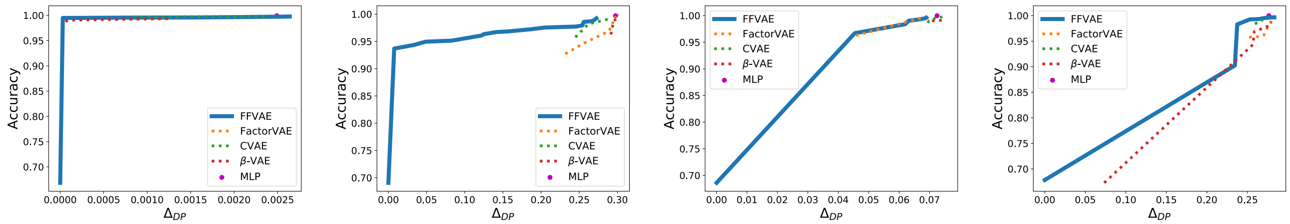
In Fig. 2b, where the sensitive attribute Shape is correlated by design with y, we see the clearest improvement of the FFVAE over the baselines, with an almost complete reduction in Δ_{DP} and very little accuracy loss. The baseline models are all unable to improve Δ_{DP} by more than about 0.05, indicating that they have not effectively disentangled the sensitive information from the label in the representation.

In Figs. 2c and 2d, we examine conjunctions of sensitive attributes, assessing FFVAE's ability to flexibly provide multi-attribute fair representations. Since Shape, but not Scale, is correlated with XPosition, the disentanglement problem is slightly easier than the one in Fig. 2b. We observe that FFVAE is still better than the baselines in this problem: in Fig. 2d, its Pareto front dominates (except at one point), and in Fig. 2c, only FactorVAE matches FFVAE's results in the high-accuracy regime, and no models do in the low Δ_{DP} regime.

Taken altogether, the fair classification audits in Fig. 2 show that our method is capable of disentangling information from multiple sensitive attributes, which enables fair downstream classification via simple and composable latent code manipulations that do not hurt accuracy too much.

Disentanglement and Predictiveness In Fig. 3, we display results from our disentanglement and predictiveness audits. As explained earlier, we train two classifiers to predict some sensitive attribute A_i: one from the represen-

[4] https://github.com/deepmind/dsprites-dataset



(a) a = Scale, y = XPosition (b) a = Shape, y = XPosition (c) a = Shape ∧ Scale, y = XPosition (d) a = Shape ∨ Scale, y = XPosition

Figure 2. Fairness-accuracy tradeoff curves, DSpritesUnfair dataset. FFVAE is our model from Eq. 10. FactorVAE is the same, but with $\alpha = 0$, and CVAE has $\gamma = 0$. β-VAE is as in Eq. 7. We sweep a range of hyperparameters and report the Pareto fronts. Optimal point is the top left hand corner — this represents perfect accuracy and fairness. MLP is a baseline classifier trained directly on the input data. For each model, encoder outputs are modified to remove information about a; see text for details.

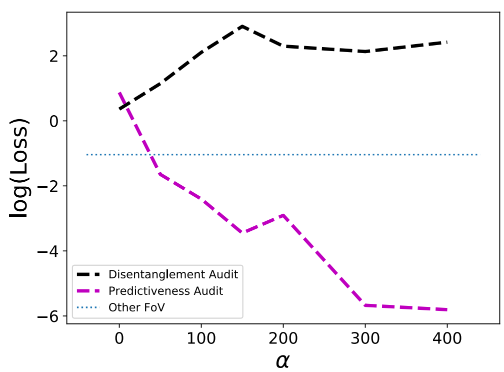


Figure 3. Auditing of FFVAE models across values of α. $\alpha = 0$ is equivalent to FactorVAE. Y-axis is the log-cross entropy of an auditing classifier model. We show an average across values of γ. Black dashed line is a disentanglement audit — training an MLP classifier to predict A_i on the learned representation with b_i (the dimension associated with Shape) removed (i.e. $[z, b_{\backslash i}]$). A high loss here means good disentanglement. Magenta dashed line is a predictiveness audit — training a logistic regression classifier to predict A_i based on only b_i. A low loss here means good predictiveness. A_i is Shape. The blue horizontal dotted line is a baseline — we train a classifier to predict A_i ignoring the image but instead using the other 5 factors of variation (FoV) in DSprites (Color, Scale, Rotation, XPosition, YPosition) as input. This represents the amount of information contained about A_i inherently in the data — we want our disentanglement audit to have higher loss than this, and our predictiveness audit to have lower loss than this line.

tation with the associated dimension b_i removed (disentanglement), and one from the associated dimension b_i only (predictiveness). The classifier loss is cross-entropy, which is a lower bound on the mutual information between the input and target of the classifier. We run these audits on each encoder that we train and plot the results in Fig. 3.

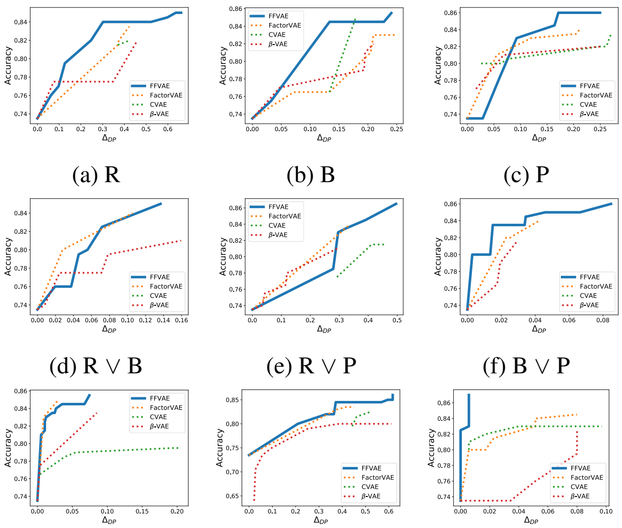
The results show that increasing α helps both predictiveness and disentanglement. In our disentanglement audit, we find that larger α makes it more difficult to predict a sensitive attribute from the rest of the representation with b_i removed. The horizontal dotted line is from a classifier which predicts a sensitive attribute from the other factors of variation in DSprites — this acts as a baseline which controls for the amount of correlation that are already present in the data. We see that when $\alpha = 0$ (i.e. FactorVAE), it is slightly more difficult than this baseline to predict the sensitive attribute. This is due to the disentanglement prior. However, increasing $\alpha > 0$ increases disentanglement benefits in FFVAE beyond what is present in FactorVAE. This shows that encouraging predictive structure can help disentanglement through isolating each attribute's information in particular latent dimensions. Additionally, we see that by increasing α we can get improved predictiveness. When $\alpha = 0$, then b_i is not a useful predictor of a_i. However, as we increase α, it becomes a virtually perfect predictor.

We further evaluate the disentanglement properties of our model in Appendix E using the Mutual Information Gap metric (Chen et al., 2018).

5.3. Communities & Crime

Dataset Communities & Crime[5] is a tabular UCI dataset containing neighborhood-level population statistics. 120 such statistics are recorded for each of the 1,994 neighborhoods. Several attributes encode demographic information that may be protected. We chose three as sensitive: racePctBlack (% neighborhood population which is Black), blackPerCap (avg per capita income of Black residents), and pctNotSpeakEnglWell (% neighborhood population that does not speak English well). We follow the same train/eval procedure as with DSpritesUnfair - we train FFVAE with the sensitive attributes and evaluate using naive MLPs to predict a held-out label (violent crimes per capita) on held-out data.

[5] http://archive.ics.uci.edu/ml/datasets/communities+and+crime



(a) R (b) B (c) P

(d) R ∨ B (e) R ∨ P (f) B ∨ P

(g) R ∧ B (h) R ∧ P (i) B ∧ P

Figure 4. Communities & Crime fair classification results. We chose the following three attributes to be sensitive: racePctBlack (R), blackPerCapIncome (B), and pctNotSpeakEnglWell (P) See text for details.

Fair Classification This dataset presents a more difficult disentanglement problem than DSpritesUnfair. The three sensitive attributes we chose in Communities and Crime were somewhat correlated with each other, a natural artefact of using real (rather than simulated) data. We note that in general, the disentanglement literature does not provide much guidance in terms of disentangling correlated attributes. However, despite this obstacle, we found that the FFVAE performed reasonably well in the fair classification audit (Fig. 4). In general, it achieved higher accuracy than the baseline methods, possibly due to its ability to incorporate supervision during training. The FactorVAE tended to be the best-performing baseline in this task, suggesting that the most important component of the model for fair classification is the independence regularization. While our method does not outperform the baselines on each conjunction, its relatively strong performance on a difficult, tabular dataset shows the promise of using disentanglement priors in designing robust subgroup-fair machine learning models.

6. Discussion

In this paper we have connected disentangled representation learning to the goals of subgroup fair machine learning, and presented a method for learning a structured latent code using multiple sensitive attributes. The proposed encoder provides *flexibly fair* representations, which can be modified simply and compositionally at test time to yield a fair representation with respect to multiple sensitive attributes and their conjunctions, even when test-time sensitive attribute labels are unavailable. Empirically we found that our encoder disentangled sensitive sources of variation in synthetic image data, even in the challenging scenario of skewed training data. Our method compared favorably with baseline disentanglement algorithms in downstream fair classifications by achieving better parity for a given accuracy budget across several group and subgroup definitions. None of the models performed robustly across all possible subgroups in the real-data setting. This result reflects the difficulty of subgroup fair representation learning and motivates further work in this area.

There are two main directions of interest for future work. First is the question of fairness metrics: a wide range of fairness metrics beyond demographic parity have been proposed (Hardt et al., 2016; Pleiss et al., 2017). Understanding how to learn flexibly fair representations with respect to other metrics would broaden the applicability of subgroup fair representation learning.

Secondly, robustness to distributional shift presents an important challenge in the context of both disentanglement and fairness. In disentanglement, we aim to learn independent factors of variation. Most empirical work on evaluating disentanglement has used synthetic data with uniformly distributed factors of variation, but this setting is unrealistic. Meanwhile, in fairness, we hope to learn from potentially biased data distributions, which may suffer from both undersampling and systemic historical discrimination. We wish to imagine hypothetical "unbiased" data or compute robustly fair representations. Therefore, fairness and disentanglement under distributional shift is a rich area for exploration. We hope that this work sheds light on these issues, and represents a first step towards understanding and leveraging the relationship between fair and disentangled representation learning.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken elbo. In *International Conference on Machine Learning*, pp. 159–168, 2018.

Aleo, M. and Svirsky, P. Foreclosure fallout: the banking industrys attack on disparate impact race discrimination claims under the fair housing act and the equal credit opportunity act. *Public Law Interest Journal*, 18(1):1–66, 2008. URL https://www.bu.edu/pilj/files/2015/09/18-1AleoandSvirskyArticle.pdf.

Ardizzone, L., Kruse, J., Wirkert, S., Rahner, D., Pellegrini, E. W., Klessen, R. S., Maier-Hein, L., Rother, C., and Köthe, U. Analyzing inverse problems with invertible neural networks. *arXiv preprint arXiv:1808.04730*, 2018.

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul): 1–48, 2002.

Barocas, S. and Selbst, A. D. Big Data's Disparate Impact. *SSRN eLibrary*, 2014.

Bose, A. J. and Hamilton, W. L. Compositional fairness constraints for graph embeddings. *Relational Representation Learning Workshop, Neural Information Processing Systems 2018*, 2018.

Botros, P. and Tomczak, J. M. Hierarchical vampprior variational fair auto-encoder. *arXiv preprint arXiv:1806.09918*, 2018.

Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pp. 77–91, 2018.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

DCCA. Division of consumer and community affairs. 2011-07. 12 cfr supplement i to part 202 - official staff interpretations. https://www.law.cornell.edu/cfr/text/12/appendix-Supplement_I_to_part_202, 1983.

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

Elliot, M. N., Fremont, A., Morrison, P. A., Pantoja, P., and Lurie, N. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack SelfReported Race/Ethnicity. *Health Services Research*, 2008.

Esmaeili, B., Wu, H., Jain, S., Bozkurt, A., Siddharth, N., Paige, B., Brooks, D. H., Dy, J., and van de Meent, J.-W. Structured disentangled representations. *stat*, 1050:29, 2018.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hebert-Johnson, U., Kim, M., Reingold, O., and Rothblum, G. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/hebert-johnson18a.html.

Hellman, D. Indirect discrimination and the duty to avoid compounding injustice. *Foundations of Indirect Discrimination Law*, 2018.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Hyvärinen, A. and Oja, E. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5): 411–430, 2000.

Jacobsen, J.-H., Behrmann, J., Zemel, R., and Bethge, M. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *arXiv preprint arXiv:1711.05144*, 2017.

Kearns, M. J., Neel, S., Roth, A., and Wu, Z. S. An empirical study of rich subgroup fairness for machine learning. *arXiv preprint arXiv:1808.08166*, 2018.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kim18b.html.

Kim, M. P., Ghorbani, A., and Zou, J. Multiaccuracy: Blackbox post-processing for fairness in classification. *arXiv preprint arXiv:1805.12317*, 2018.

Kim, S.-E., Paik, H. Y., Yoon, H., Lee, J. E., Kim, N., and Sung, M.-K. Sex- and gender-specific disparities in colorectal cancer risk. *World Journal of Gastroentorology*, 21(17):5167–5175, 2015.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Kirchner, L., Mattu, S., Larson, J., and Angwin, J. Machine Bias: Theres Software Used Across the Country to Predict Future Criminals. And its Biased Against Blacks., May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Klys, J., Snell, J., and Zemel, R. Learning latent subspaces in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 6443–6453, 2018.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 4066–4076. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6995-counterfactual-fairness.pdf.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018.

Moyer, D., Gao, S., Brekelmans, R., Galstyan, A., and Ver Steeg, G. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pp. 9101–9110, 2018.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5680–5689. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7151-on-fairness-and-calibration.pdf.

Rothenhusler, D., Meinshausen, N., Bhlmann, P., and Peters, J. Anchor regression: heterogeneous data meets causality. *arXiv preprint arXiv:1801.06229*, 2018.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. *arXiv preprint arXiv:1812.04218*, 2018.

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