Learning Controllable Fair Representations

Jiaming Song TSONG@CS.STANFORD.EDU
Pratyusha Kalluri PKALLURI@STANFORD.EDU
Aditya Grover ADITYAG@CS.STANFORD.EDU
Shengjia Zhao SJZHAO@STANFORD.EDU
Stefano Ermon ERMON@CS.STANFORD.EDU
Computer Science Department
Stanford University
Stanford, CA 94305-2018, USA

Abstract

Learning data representations that are transferable and fair with respect to certain protected attributes is crucial to reducing unfair decisions made downstream, while preserving the utility of the data. We propose an information-theoretically motivated objective for learning maximally expressive representations subject to fairness constraints. We demonstrate that a range of existing approaches optimize approximations to the Lagrangian dual of our objective. In contrast to these existing approaches, our objective provides the user control over the fairness of representations by specifying limits on unfairness. We introduce a dual optimization method that optimizes the model as well as the expressiveness-fairness trade-off. Empirical evidence suggests that our proposed method can account for multiple notions of fairness and achieves higher expressiveness at a lower computational cost.

1. INTRODUCTION

Statistical learning systems are increasingly being used to assess individuals, influencing consequential decisions such as bank loans, college admissions, and criminal sentences. This yields a growing demand for systems guaranteed to output decisions that are fair with respect to sensitive attributes such as gender, race, and disability.

In typical classification and regression settings with fairness and privacy constraints, one is usually concerned about a single, specific task. However, situations might arise where a data owner needs to release data to downstream users without prior knowledge of the tasks that will be performed (Madras et al., 2018). In such cases, it is crucial to find representations of the data that can be used on a wide variety of tasks while preserving fairness (Calmon et al., 2017).

This gives rise to two desiderata. On one hand, the representations need to be *expressive*, so that they can be used effectively for as many tasks as possible. On the other hand, the representations also need to satisfy certain *fairness* constraints to protect sensitive attributes. However, many notions of fairness are possible, and it might not be possible to simultaneously satisfy all of them (Kleinberg et al., 2016; Chouldechova, 2017). Therefore, the ability to effectively trade off multiple notions of fairness is crucial to fair representation learning.

To this end, we present a constrained optimization framework under an information-theoretic perspective (Section 2). The goal, learning controllable fair representations, is to maximize the expressiveness of the representation while satisfying certain fairness constraints. We represent expressiveness as well as three dominant notions of fairness (demographic parity (Zemel et al., 2013), equalized odds, equalized opportunity (Hardt et al., 2016)) in terms of mutual information, obtain tractable upper/lower bounds of these mutual information objectives, and connect them with existing objectives such as maximum likelihood, adversarial training (Goodfellow et al., 2014), and variational autoencoders (Kingma and Welling, 2013; Rezende and Mohamed, 2015).

As we demonstrate in Section 3, this serves as a unifying framework for existing work (Zemel et al., 2013; Louizos et al., 2015; Edwards and Storkey, 2015; Madras et al., 2018) on learning fair representations. A range of existing approaches to learning fair representations (which do not draw connections to information theory) optimize an approximation of the Lagrangian dual of our objective with fixed values of the Lagrange multipliers. This would require the user to obtain different representations for different notions of fairness as in Madras et al. (2018).

Instead, we consider a dual optimization approach (Section 4), in which we optimize the Lagrange multipliers during training (Zhao et al., 2018), thereby also learning the trade-off between expressiveness and fairness. We further show that our proposed framework is strongly convex in distribution space.

Our work is the first to provide direct user control over the fairness of representations through fairness constraints that are easily interpretable by non-expert users. Empirical results in Section 5 demonstrate that our notions of expressiveness and fairness based on mutual information align well with existing definitions, our method encourages representations that satisfy the fairness constraints while being more expressive, and that our method is able to balance the trade-off between multiple notions of fairness with a single representation with a significantly smaller computational cost.

2. AN INFORMATION-THEORETIC OBJECTIVE FOR CONTROLLABLE FAIR REPRESENTATIONS

We are given a dataset $\mathcal{D}_u = \{(\mathbf{x}_i, \mathbf{u}_i)\}_{i=1}^M$ containing pairs of observations $\mathbf{x} \in \mathcal{X}$ and sensitive attributes $\mathbf{u} \in \mathcal{U}$. We assume the dataset is sampled i.i.d. from an unknown data distribution $q(\mathbf{x}, \mathbf{u})$. Our goal is to transform each data point $(\mathbf{x}, \mathbf{u})$ into a new *representation* $\mathbf{z} \in \mathcal{Z}$ that is (1) *transferable*, i.e., it can be used in place of $(\mathbf{x}, \mathbf{u})$ by multiple unknown vendors on a variety of downstream tasks, and (2) *fair*, i.e., the sensitive attributes $\mathbf{u}$ are protected. For conciseness, we focus on the *demographic parity* notion of fairness (Calders et al., 2009; Zliobaite, 2015; Zafar et al., 2015), which requires the decisions made by a classifier over $\mathbf{z}$ to be independent of the sensitive attributes $\mathbf{u}$. We discuss in Appendix C how our approach can be extended to control other notions of fairness simultaneously, such as the *equalized odds* and *equalized opportunity* notions of fairness (Hardt et al., 2016).

We assume the representations $\mathbf{z} \in \mathcal{Z}$ for $(\mathbf{x}, \mathbf{u})$ are obtained by sampling from a conditional probability distribution $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$ parameterized by $\phi \in \Phi$. The joint distribution of $(\mathbf{x}, \mathbf{z}, \mathbf{u})$ is then given by $q_\phi(\mathbf{x}, \mathbf{z}, \mathbf{u}) = q(\mathbf{x}, \mathbf{u}) q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$. We formally express our desiderata for learning a controllable fair representation $\mathbf{z}$ through the concept of mutual information:

1. **Fairness**: $\mathbf{z}$ should have low mutual information with the sensitive attributes $\mathbf{u}$.
2. **Expressiveness**: $\mathbf{z}$ should have high mutual information with the observations $\mathbf{x}$ conditioned on $\mathbf{u}$ (in expectation over possible values of $\mathbf{u}$).

The first condition encourages $\mathbf{z}$ to be independent of $\mathbf{u}$; if this is indeed the case, the downstream vendor cannot learn a classifier over the representations $\mathbf{z}$ that discriminates based on $\mathbf{u}$. Intuitively, the mutual information $I_q(\mathbf{z}, \mathbf{u})$ is related to the optimal predictor of $\mathbf{u}$ given $\mathbf{z}$. If $I_q(\mathbf{z}, \mathbf{u})$ is zero, then no classifier can perform better than chance; if $I_q(\mathbf{z}, \mathbf{u})$ is large, vendors in downstream tasks could utilize $\mathbf{z}$ to predict the sensitive attributes $\mathbf{u}$ and make unfair decisions.

The second condition encourages $\mathbf{z}$ to contain as much information as possible from $\mathbf{x}$ conditioned on the knowledge of $\mathbf{u}$. By conditioning on $\mathbf{u}$, we ensure we do not encourage information in $\mathbf{x}$ that is correlated with $\mathbf{u}$ to leak into $\mathbf{z}$. In general, the two desiderata allow $\mathbf{z}$ to encode non-sensitive information from $\mathbf{x}$ (expressiveness) while excluding information in $\mathbf{u}$ (fairness).

Our goal is to choose parameters $\phi \in \Phi$ for $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})$ that meets both these criteria[1]. Because we wish to ensure our representations satisfy fairness constraints even at the cost of using less expressive $\mathbf{z}$, we synthesize the two desiderata into the following constrained optimization problem:

$$\max_{\phi \in \Phi} I_q(\mathbf{x};\mathbf{z}|\mathbf{u}) \qquad \text{s.t. } I_q(\mathbf{z};\mathbf{u}) < \epsilon \tag{1}$$

where $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$ denotes the mutual information of $\mathbf{x}$ and $\mathbf{z}$ conditioned on $\mathbf{u}$, $I_q(\mathbf{z};\mathbf{u})$ denotes mutual information between $\mathbf{z}$ and $\mathbf{u}$, and the hyperparameter $\epsilon > 0$ controls the maximum amount of mutual information allowed between $\mathbf{z}$ and $\mathbf{u}$. The motivation of our "hard" constraint on $I_q(\mathbf{z};\mathbf{u})$ – as opposed to a "soft" regularization term – is that even at the cost of learning less expressive $\mathbf{z}$ and losing some predictive power, we view as important ensuring that our representations are fair to the extent dictated by ϵ.

Both mutual information terms in Equation 1 are difficult to compute and optimize. In particular, the optimization objective in Equation 1 can be expressed as the following expectation:

$$\begin{aligned}&I_q(\mathbf{x};\mathbf{z}|\mathbf{u})\\ &= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{x},\mathbf{z}|\mathbf{u}) - \log q(\mathbf{x}|\mathbf{u}) - \log q_\phi(\mathbf{z}|\mathbf{u})]\end{aligned}$$

while the constraint on $I_q(\mathbf{z};\mathbf{u})$ involves the following expectation:

$$I_q(\mathbf{z};\mathbf{u}) = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{z}|\mathbf{u}) - \log q_\phi(\mathbf{z})]$$

Even though $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})$ is known analytically (and assumed to be easy to evaluate), both mutual information terms are difficult to estimate and optimize.

To offset the challenge in estimating mutual information, we introduce upper and lower bounds with tractable Monte Carlo gradient estimates. We introduce the following lemmas, with the proofs provided in Appendix A. We note that similar bounds have been proposed in Alemi et al. (2016, 2017); Zhao et al. (2018).

2.1 Tractable Lower Bound for $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$

We begin with a (variational) *lower* bound on the objective function $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$ related to expressiveness which we would like to *maximize* in Equation 1.

Lemma 1 *For any conditional distribution $p(\mathbf{x}|\mathbf{z},\mathbf{u})$*

$$\begin{aligned}I_q(\mathbf{x};\mathbf{z}|\mathbf{u}) = &\ \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})] + H_q(\mathbf{x}|\mathbf{u})\\ &+ \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{x}|\mathbf{z},\mathbf{u}) \| p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u}))\end{aligned}$$

where $H_q(\mathbf{x}|\mathbf{u})$ is the entropy of $\mathbf{x}$ conditioned on $\mathbf{u}$, and D_{KL} denotes KL-divergence.

Since entropy and KL divergence are non-negative, the above lemma implies the following lower bound:

$$I_q(\mathbf{x};\mathbf{z}|\mathbf{u}) \geq \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})] := \mathcal{L}_r. \tag{2}$$

2.2 Tractable Upper Bound for $I_q(\mathbf{z};\mathbf{u})$

Next, we provide an *upper* bound for the constraint term $I_q(\mathbf{z};\mathbf{u})$ that specifies the limit on unfairness. In order to satisfy this fairness constraint, we wish to implicitly *minimize* this term.

1. Simply ignoring $\mathbf{u}$ as an input is insufficient, as $\mathbf{x}$ may still contain information about $\mathbf{u}$.

Lemma 2 *For any distribution $p(\mathbf{z})$, we have:*

$$\begin{aligned} I_q(\mathbf{z};\mathbf{u}) &\leq I_q(\mathbf{z};\mathbf{x},\mathbf{u}) \\ &= \mathbb{E}_{q(\mathbf{x},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) \| p(\mathbf{z})) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}) \| p(\mathbf{z})). \end{aligned} \tag{3}$$

Again, using the non-negativity of KL divergence, we obtain the following upper bound:

$$I_q(\mathbf{z};\mathbf{u}) \leq \mathbb{E}_{q(\mathbf{x},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) \| p(\mathbf{z})) := C_1. \tag{4}$$

In summary, Equation 2 and Equation 4 imply that we can compute tractable Monte Carlo estimates for the lower and upper bounds to $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$ and $I_q(\mathbf{z};\mathbf{u})$ respectively, as long as the variational distributions $p(\mathbf{x}|\mathbf{z},\mathbf{u})$ and $p(\mathbf{z})$ can be evaluated tractably, e.g., Bernoulli and Gaussian distributions. Note that the distribution $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})$ is assumed to be tractable.

2.3 A Tighter Upper Bound to $I_q(\mathbf{z},\mathbf{u})$ via Adversarial Training

It would be tempting to use C_1, the tractable upper bound from Equation 4, as a replacement for $I_q(\mathbf{z},\mathbf{u})$ in the constraint of Equation 1. However, note from Equation 3 that C_1 is *also* an upper bound to $I_q(\mathbf{x},\mathbf{z}|\mathbf{u})$, which is the objective function (expressiveness) we would like to maximize in Equation 1. If the constraint is too tight, we would constrain the expressiveness of our learned representations. Therefore, we introduce a tighter bound via the following lemma.

Lemma 3 *For any distribution $p(\mathbf{u})$, we have:*

$$I_q(\mathbf{z};\mathbf{u}) = \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) - D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u})). \tag{5}$$

Using the non-negativity of KL divergence as before, we obtain the following upper bound on $I_q(\mathbf{z};\mathbf{u})$:

$$I_q(\mathbf{z};\mathbf{u}) \leq \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) := \widehat{C}_2. \tag{6}$$

As $\mathbf{u}$ is typically low-dimensional (e.g., a binary variable, as in Hardt et al. (2016); Zemel et al. (2013)), we can choose $p(\mathbf{u})$ in Equation 5 to be a kernel density estimate based on the dataset $\mathcal{D}$. By making $D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u}))$ as small as possible, our upper bound $\widehat{C}_2$ gets closer to $I_q(\mathbf{z},\mathbf{u})$.

While $\widehat{C}_2$ is a valid upper bound to $I_q(\mathbf{z};\mathbf{u})$, the term $q_\phi(\mathbf{u}|\mathbf{z})$ appearing in $\widehat{C}_2$ is intractable to evaluate, requiring an integration over $\mathbf{x}$. Our solution is to approximate $q_\phi(\mathbf{u}|\mathbf{z})$ with a parametrized model $p_\psi(\mathbf{u}|\mathbf{z})$ with parameters $\psi \in \Psi$ obtained via the following objective:

$$\min_\psi \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p_\psi(\mathbf{u}|\mathbf{z})). \tag{7}$$

Note that the above objective corresponds to maximum likelihood prediction with inputs $\mathbf{z}$ and labels $\mathbf{u}$ using $p_\psi(\mathbf{u}|\mathbf{z})$. In contrast to $q_\phi(\mathbf{u}|\mathbf{z})$, the distribution $p_\psi(\mathbf{u}|\mathbf{z})$ is tractable and implies the following lower bound to $\widehat{C}_2$:

$$\begin{aligned} &\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})] \\ &= \mathbb{E}_{q_\phi(\mathbf{z})}[D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) - D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p_\psi(\mathbf{u}|\mathbf{z}))] \\ &\leq \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) = \widehat{C}_2. \end{aligned}$$

It follows that we can approximate $I_q(\mathbf{z};\mathbf{u})$ through the following adversarial training objective:

$$\min_\phi \max_\psi \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})] \tag{8}$$

Here, the goal of the adversary p_ψ is to minimize the difference between the tractable approximation given by $\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})]$ and the intractable true upper bound $\widehat{C}_2$. We summarize this observation in the following result:

Corollary 4 *If* $D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z})\|p_\psi(\mathbf{u}|\mathbf{z})) \leq \ell$, *then*

$$\begin{aligned}I_q(\mathbf{z};\mathbf{u}) \leq\; &\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})] \\ &- D_{\mathrm{KL}}(q(\mathbf{u})\|p(\mathbf{u})) + \ell\end{aligned}$$

for any distribution $p(\mathbf{u})$.

It immediately follows that when $\ell \to 0$, i.e., the adversary approaches global optimality, we obtain the true upper bound. For any other finite value of ℓ, we have:

$$\begin{aligned}I_q(\mathbf{z};\mathbf{u}) &\leq \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})] + \ell \\ &:= C_2 + \ell. \end{aligned} \tag{9}$$

2.4 A practical objective for controllable fair representations

Recall that our goal is to find tractable estimates to the mutual information terms in Equation 1 to make the objective and constraints tractable. In the previous sections, we have derived a lower bound for $I_q(\mathbf{x},\mathbf{u}|\mathbf{z})$ (which we want to maximize) and upper bounds for $I_q(\mathbf{u},\mathbf{z})$ (which we want to implicitly minimize to satisfy the constraint). Therefore, by applying these results to the optimization problem in Equation 1, we obtain the following constrained optimization problem:

$$\begin{aligned}\min_{\theta,\phi}\max_{\psi\in\Psi}\quad &\mathcal{L}_r = -\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})] \\ \text{s.t.}\quad &C_1 = \mathbb{E}_{q(\mathbf{x},\mathbf{u})}D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})\|p(\mathbf{z})) < \epsilon_1 \\ &C_2 = \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log p_\psi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u})] < \epsilon_2\end{aligned} \tag{10}$$

where $\mathcal{L}_r$, C_1, and C_2 are introduced in Equations 2, 4 and 6 respectively.

Both C_1 and C_2 provide a way to limit $I_q(\mathbf{z},\mathbf{u})$. C_1 is guaranteed to be an upper bound to $I_q(\mathbf{z};\mathbf{u})$ but also upper-bounds $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$ (which we would like to maximize), so it is more suitable when we value true guarantees on fairness over expressiveness. C_2 may more accurately approximate $I_q(\mathbf{z};\mathbf{u})$ but is guaranteed to be an upper bound only in the case of an optimal adversary. Hence, it is more suited for scenarios where the user is satisfied with guarantees on fairness in the limit of adversarial training, and we wish to learn more expressive representations. Depending on the underlying application, the user can effectively remove either of the constraints C_1 or C_2 (or even both) by setting the corresponding ϵ to infinity.

3. A UNIFYING FRAMEWORK FOR RELATED WORK

Multiple methods for learning fair representations have been proposed in the literature. Zemel et al. (2013) propose a method for clustering individuals into a small number of discrete fair representations. Discrete representations, however, lack the representational power of distributed representations, which vendors desire. In order to learn distributed fair representations, Edwards and Storkey (2015) and Madras et al. (2018) each propose adversarial training, where the latter (LAFTR) connects different adversarial losses to multiple notions of fairness. Louizos et al. (2015) propose VFAE for learning distributed fair representations by using a variational autoencoder architecture with additional regularization based on Maximum Mean Discrepancy (MMD) (Gretton et al., 2007). Each of these methods is limited to the case of a binary sensitive attribute because their measurements of fairness are based on statistical parity (Zemel et al., 2013), which is defined only for two groups.

Moreover, each of these methods can be viewed as optimizing an *approximation* of the Lagrangian dual of our objective in Equation 10, with particular *fixed* settings of the Lagrangian multipliers:

$$\begin{aligned}&\arg\min_{\theta,\phi}\max_{\psi}\mathcal{L}_r + \lambda_1(C_1 - \epsilon_1) + \lambda_2(C_2 - \epsilon_2) \\ =&\arg\min_{\theta,\phi}\max_{\psi}\mathcal{L}_r + \lambda_1 C_1 + \lambda_2 C_2\end{aligned} \tag{11}$$

where $\mathcal{L}_r$ and C_i are defined as in Equation 10, ϵ_i specify the set of feasible solutions, and the multipliers $\lambda_i \geq 0$ are hyperparameters controlling the relative strengths of the constraints (which now act as "soft" regularizers).

We use "approximation" to suggest these objectives are not exactly the same as ours, as ours can deal with more than two groups in the fairness criterion C_2 and theirs cannot. However, all the fairness criteria achieve $\mathbf{z} \perp \mathbf{u}$ at a global optimum; in the following discussions, for brevity we use C_2 to indicate their objectives, even when they are not identical to ours[2].

Here, the values of ϵ do not affect the final solution. Therefore, if we wish to find representations that satisfy specific constraints, we would have to search over the hyperparameter space to find feasible solutions, which could be computationally inefficient. We call this class of approaches *Mutual Information-based Fair Representations* (MIFR). In Table 1, we summarize these existing methods.

Table 1: Summarizing the components in existing methods. The hyperparameters (e.g. A_z, α, β) are from the original notations of the corresponding methods.

	λ_1	λ_2
Zemel et al. (2013)	0	A_z/A_x
Edwards and Storkey (2015)	0	α/β
Madras et al. (2018)	0	γ/β
Louizos et al. (2015)	1	β

- Zemel et al. (2013) consider $\mathcal{L}_r$ as well as minimizing statistical parity (Equation 4 in their paper); they assume $\mathbf{z}$ is discrete, bypassing the need for adversarial training. Their objective is equivalent to Equation 11 with $\lambda_1 = 0, \lambda_2 = A_z/A_x$.

- Edwards and Storkey (2015) considers $\mathcal{L}_r$ (where $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u})$ is Gaussian) and adversarial training where the adversary tries to distinguish the representations from two groups (Equation 9). Their objective is equivalent to Equation 11 with $\lambda_1 = 0, \lambda_2 = \alpha/\beta$.

- Madras et al. (2018) considers $\mathcal{L}_r$ and adversarial training, which optimizes over surrogates to the demographic parity distance between two groups (Equation 4). Their objective is equivalent to Equation 11 with $\lambda_1 = 0, \lambda_2 = \gamma/\beta$.

- Louizos et al. (2015) consider $\mathcal{L}_r$, C_1 with $\lambda_1 = 1$ and the maximum mean discrepancy between two sensitive groups (C_2) (Equation 8). However, as $\mathcal{L}_r + C_1$ is the VAE objective, their solutions does not prefer high mutual information between $\mathbf{x}$ and $\mathbf{z}$ (referred to as the "information preference" property (Chen et al., 2016; Zhao et al., 2017b,a, 2018)). Their objective is equivalent to Equation 11 with $\lambda_1 = 1, \lambda_2 = \beta$.

All of the above methods requires hand-tuning λ to govern the trade-off between the desiderata (except for $\lambda_1 = 1$ in Louizos et al. (2015)). Because each of these approaches optimizes the dual with *fixed* multipliers instead of *optimizing* the multipliers to satisfy the fairness constraints, ϵ is ignored, so these approaches cannot ensure that the fairness constraints are satisfied. Using any of these approaches to empirically achieve a desirable limit on unfairness requires manually tuning the multipliers (e.g., increase some λ_i until the corresponding constraint is satisfied) over many experiments and is additionally difficult because there is no interpretable relationship between the multipliers and a limit on unfairness.

2. We also have not included the classification error in their methods, as we do not assume access to labels in our setting.

4. DUAL OPTIMIZATION FOR CONTROLLABLE FAIR REPRESENTATIONS

In order to exactly solve the dual of our practical objective from Equation 10 and guarantee that the fairness constraints are satisfied, we must optimize the model parameters as well as the Lagrangian multipliers, e.g. using the following dual objective:

$$\max_{\boldsymbol{\lambda} \geq 0} \min_{\theta,\phi} \max_{\psi} \mathcal{L} = \mathcal{L}_r + \boldsymbol{\lambda}^\top (\mathbf{C} - \boldsymbol{\epsilon}) \tag{12}$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]$ are the multipliers and $\boldsymbol{\epsilon} = [\epsilon_1, \epsilon_2]$ and $\mathbf{C} = [C_1, C_2]$ represent the constraints.

If we assume we are optimizing in the distribution space (i.e. Φ, Θ corresponds to the set of all valid distributions $(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}), p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u}), p_\theta(\mathbf{z})))$, then we can show that strong duality holds (our primal objective from Equation 10 equals our dual objective from Equation 12).

Theorem 5 *If $\epsilon_1, \epsilon_2 > 0$, then strong duality holds for the following optimization problem over distributions p_θ and q_ϕ:*

$$\min_{p_\theta, q_\phi} \quad -\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u})] \tag{13}$$

$$\begin{aligned} s.t. \quad & \mathbb{E}_{q(\mathbf{x},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_\theta(\mathbf{z})) < \epsilon_1 \\ & \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) < \epsilon_2 \end{aligned} \tag{14}$$

where q_ϕ denotes $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$ and p_θ denotes $p_\theta(\mathbf{z})$ and $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u})$.

We show the complete proof in Appendix A.4. Intuitively, we utilize the convexity of KL divergence (over the pair of distributions) and mutual information (over the conditional distribution) to verify that Slater's conditions hold for this problem. The techniques are similar to that in (Zhao et al., 2018), yet we extend it to the case with sensitive attributes $\mathbf{u}$ and $I_q(\mathbf{z}, \mathbf{u})$ in the constraints.

In practice, we can perform standard iterative gradient updates in the parameter space: standard gradient descent over θ, ϕ, gradient ascent over ψ (which parameterizes only the adversary), and gradient ascent over $\boldsymbol{\lambda}$. Intuitively, the gradient ascent over $\boldsymbol{\lambda}$ corresponds to a multiplier $\boldsymbol{\lambda}$ increasing when its constraint is not being satisfied, encouraging the representations to satisfy the fairness constraints even at a cost to representation expressiveness. Empirically, we show that this scheme is effective despite non-convexity in the parameter space.

Note that given finite model capacity, an $\boldsymbol{\epsilon}$ that is too small may correspond to no feasible solutions in the parameter space; that is, it may be impossible for the model to satisfy the specified fairness constraints. Here we introduce heuristics to estimate the mimimum feasible $\boldsymbol{\epsilon}$. The minimum feasible ϵ_1 and ϵ_3 can be estimated by running the standard conditional VAE algorithm on the same model and estimating the value of each divergence. Feasible ϵ_2 can be approximated by $H_q(\mathbf{u})$, since $I_q(\mathbf{z}; \mathbf{u}) \leq H_q(\mathbf{u})$; This can easily be estimated empirically when $\mathbf{u}$ is binary or discrete.

5. EXPERIMENTS

5.1 Experimental Setup

We evaluate our results on three datasets (Zemel et al., 2013; Louizos et al., 2017; Madras et al., 2018). The first is the UCI *German* credit dataset[3], which contains information about 1000 individuals, with a binary sensitive feature being whether the individual's age exceeds a threshold. The downstream task is to predict whether the individual is offered credit or not. The second is the UCI *Adult* dataset[4],

3. https://archive.ics.uci.edu/ml/datasets
4. https://archive.ics.uci.edu/ml/datasets/adult

which contains information of over 40,000 adults from the 1994 US Census. The downstream task is to predict whether an individual earns more than $50K/year. We consider the sensitive attribute to be gender, which is pre-processed to be a binary value. The third is the Heritage *Health* dataset[5], which contains information of over 60,000 patients. The downstream task is to predict whether the Charlson Index (an estimation of patient mortality) is greater than zero. Diverging from previous work (Madras et al., 2018), we consider sensitive attributes to be age and gender, where there are 9 possible age values and 2 possible gender values; hence the sensitive attributes have 18 configurations. This prevents VFAE (Louizos et al., 2015) and LAFTR (Madras et al., 2018) from being applied, as both methods reply on some statistical distance between two groups, which is not defined when there are 18 groups in question[6].

We assume that the model does not have access to labels during training; instead, it supplies its representations to a (unknown vendor's) classifier, whose task is to achieve high prediction on the downstream task (with labels). We compare the performance of MIFR, our model with fixed multipliers, and L-MIFR, our model using the Lagrangian dual optimization method. We provide details of the experimental setup in Appendix B. Specifically, we consider the simpler form for $p(\mathbf{z})$ commonly used in VAEs, where $p(\mathbf{z})$ is a fixed prior; the use of other more flexible parametrized forms of $p(\mathbf{z})$, such as normalizing flows (Dinh et al., 2016; Rezende and Mohamed, 2015) and autoregressive models (Kingma et al., 2016; van den Oord et al., 2016), is left as future work.

We estimate the mutual information values $I_q(\mathbf{x}; \mathbf{z}|\mathbf{u})$ and $I_q(\mathbf{u}; \mathbf{z})$ on the test set using the following equations:

$$I_q(\mathbf{x}; \mathbf{z}|\mathbf{u}) = \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) - \log q_\phi(\mathbf{z}|\mathbf{u})]$$
$$I_q(\mathbf{u}; \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{z}|\mathbf{u}) - \log q(\mathbf{u})]$$

where $q(\mathbf{u})$ is estimated via the empirical statistics over the training set, and $q_\phi(\mathbf{z}|\mathbf{u})$ is estimated via kernel density estimation over samples from $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$ with $(\mathbf{x}, \mathbf{u})$ sampled from the training set. Kernel density estimates are reasonable since both $\mathbf{z}$ and $\mathbf{u}$ are low dimensional (for example, *Adult* considers a 10-dimension $\mathbf{z}$ for 40,000 individuals). However, computing $q(\mathbf{u})$ or $q_\phi(\mathbf{z}|\mathbf{u})$ requires a summation over the training set, so we only compute these mutual information quantities during evaluation.

We aim to answer the following questions:

1. Do our information-theoretical objectives align well with existing notions of fairness?
2. Can our constraints achieve their intended effect?
3. How does L-MIFR compare with MIFR when solving controllable fair representations?
4. How are the learned representations affected by other hyperparameters, such as the number of iterations used for adversarial training (C_2)?
5. Does L-MIFR have the potential to balance different notions of fairness?

5.2 Relationship between Mutual Information, Prediction Accuracy, and Fairness Criteria

We investigate the relationship between mutual information and prediction performance by considering area under the ROC curve (AUC) for prediction tasks. We also investigate the relationship between mutual information and traditional fairness metrics by considering the Δ_{DP} fairness metric in Madras et al. (2018), which compares the absolute expected difference in classifier outcomes between two groups. Δ_{DP} is only defined on two groups of classifier outcomes, so it is not defined for the *Health*

5. https://www.kaggle.com/c/hhp
6. Δ_{DP} is only defined for binary sensitive variables in (Madras et al., 2018).

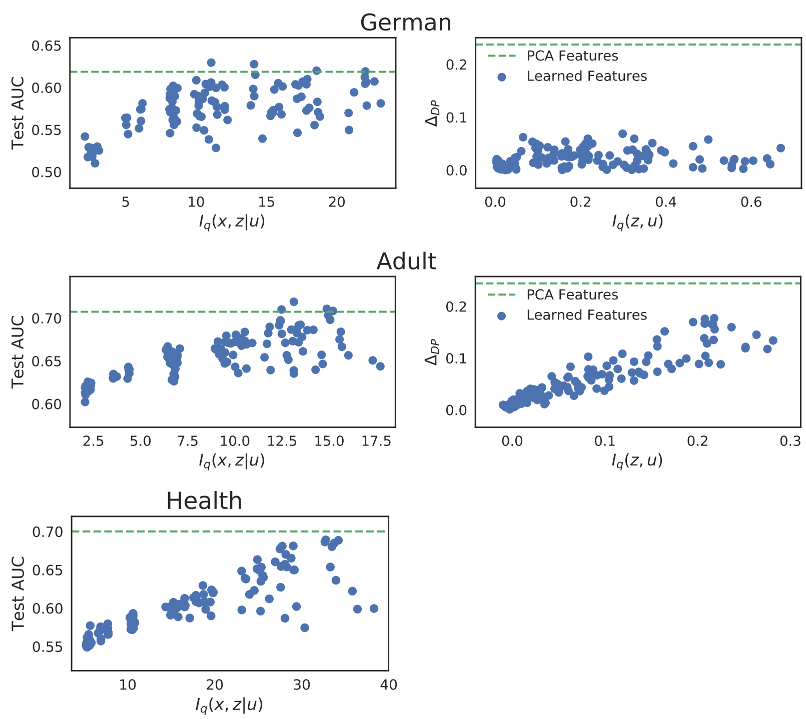

Figure 1: The relationship between mutual information and fairness related quantities. Each dot is an instance of MIFR with a different set of hyperparameters. Green line represents features obtained via principle component analysis. Increased mutual information between inputs and representations increase task performance (left) and unfairness (right).

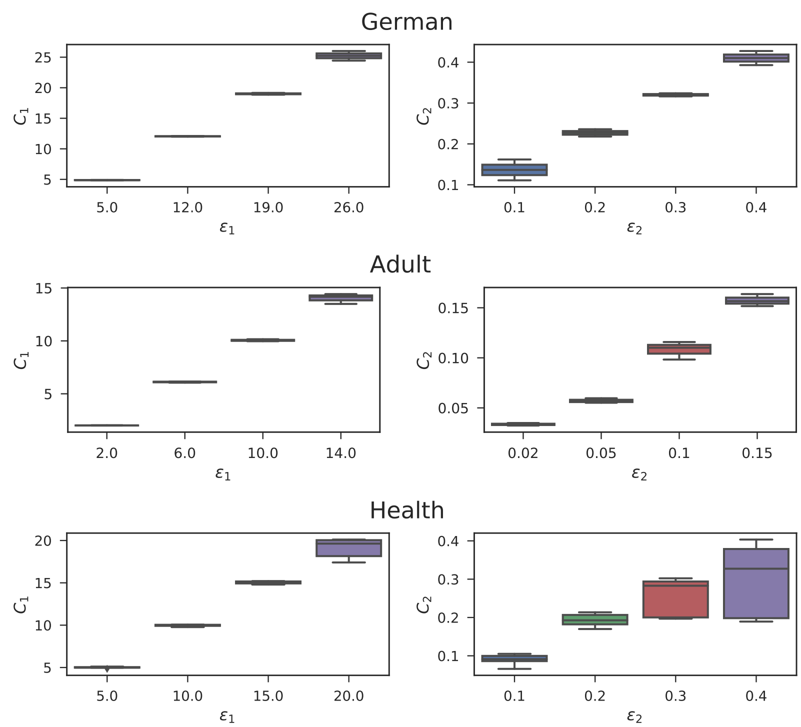

Figure 2: Corresponding C_i values under different ϵ_i with L-MIFR. After ϵ_i is fixed, we consider a range of values for the other constraint, leading to a distribution of C_i for each ϵ_i (hence the box plot).

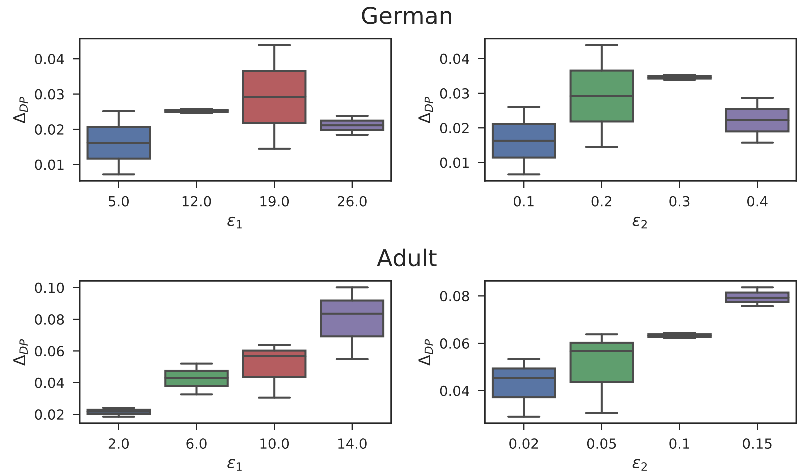

Figure 3: Δ_{DP} under different levels of ϵ with L-MIFR. Δ_{DP} generally increases as ϵ increases.

dataset when considering the sensitive attributes to be "age and gender", which has 18 groups. We use logistic regression classifiers for prediction tasks.

From the results results in Figure 1, we show that there are strong positive correlations between $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$ and test AUC, and between $I_q(\mathbf{z},\mathbf{u})$ and Δ_{DP}; increases in $I_q(\mathbf{z},\mathbf{u})$ decrease fairness. We also include a baseline in Figure 1 where the features are obtained via the top-k principle components (k is the dimension of $\mathbf{z}$), which has slightly better AUC but significantly worse fairness as measured by Δ_{DP}. Therefore, our information theoretic notions of fairness/expressiveness align well with existing notions such as Δ_{DP}/test AUC.

5.3 Controlling Representation Fairness with L-MIFR

Keeping all other constraint budgets fixed, any increase in ϵ_i for an arbitrary constraint C_i implies an increase in the unfairness budget; consequently, we are able to trade-off fairness for more informative representations.

We demonstrate this empirically via an experiment where we note the C_i values corresponding to a range of budgets ϵ_i for a fixed configuration of the other constraint budgets ϵ_j $(j \neq i)$. Then, we repeat this analysis for ϵ_j. From Figure 2, C_i increases as ϵ_i increases, and $C_i < \epsilon_i$ holds under different values of constraints ϵ_j. This suggest that we can use ϵ_i to control the solution in terms of C_i (fairness criteria).

We further show the changes in Δ_{DP} values as we vary ϵ_i in Figure 3. In *Adult*, Δ_{DP} clearly increases as ϵ_i increases; this is less obvious in *German*, as Δ_{DP} is already very low. These results suggest that the user can control the level of fairness for the representations with L-MIFR quantitatively via ϵ.

5.4 Improving Representation Expressiveness with L-MIFR

Recall that our goal is to perform controlled fair representation learning, which requires us to learn expressive representations subject to fairness constraints. We compare two approaches that could achieve this: 1) MIFR, which has to consider a range of multipliers (e.g. from grid search) to obtain solutions that satisfy the constraints; 2) L-MIFR, which finds feasible solutions directly by optimizing the Lagrange multipliers.

We evaluate both methods on 4 sets of constraints by modifying the values of ϵ_2 (which is the tighter estimate of $I_q(\mathbf{z};\mathbf{u})$) while keeping ϵ_1 fixed, and compare the expressiveness of the

Dataset	D	1	2	5	10
Adult	$I_q(\mathbf{x};\mathbf{z}\|\mathbf{u})$	10.46	10.94	9.75	9.54
	$I_q(\mathbf{z};\mathbf{u})$	0.10	0.07	0.08	0.06
Health	$I_q(\mathbf{x};\mathbf{z}\|\mathbf{u})$	16.60	16.47	16.65	16.75
	$I_q(\mathbf{z};\mathbf{u})$	0.17	0.17	0.22	0.28

Table 2: Expressiveness and fairness of the representations from L-MIFR under various D.

features learned by the two methods in Figure 4. For MIFR, we perform a grid search over $5^2 = 25$ configurations. We run one instance of L-MIFR for each ϵ setting, which takes roughly the same time to run one instance of MIFR (the only overhead is updating the two scalar values λ_1 and λ_2).

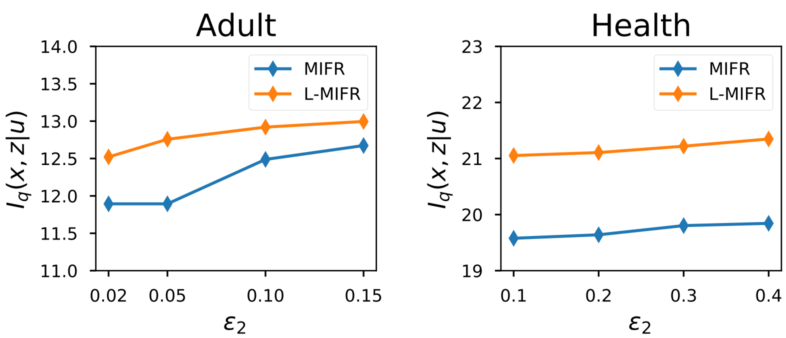

Figure 4: Expressiveness vs. ϵ_2. A larger feasible region (as measured by ϵ_2) leads to more expressive representations (as measured by $I_q(\mathbf{x}, \mathbf{z}|\mathbf{u})$).

In terms of representation expressiveness, L-MIFR outperforms MIFR even though MIFR took almost 25x the computational resources. Therefore, L-MIFR is much more computationally efficient than MIFR in terms of controlled fair representation learning.

5.5 Ablation Studies

The C_2 objective requires adversarial training, which involves iterative training of (θ, ϕ) with with ψ. In this section, we discuss the sensitivity of the expressiveness and fairness of the learned representations to the number of iterations D for ψ per iteration for (θ, ϕ). Following practices in (Gulrajani et al., 2017) to have more iterations for critic, we consider $D = \{1, 2, 5, 10\}$, and use the same number of total iterations for training.

In Table 2, we evaluate $I_q(\mathbf{x}; \mathbf{z}|\mathbf{u})$ and $I_q(\mathbf{z}; \mathbf{u})$ obtained L-MIFR on *Adult* ($\epsilon_2 = 0.10$) and *Health* ($\epsilon_2 = 0.30$). This suggests that the final solution of the representations is not very sensitive to D, although larger D seem to find solutions that are closer to ϵ_2.

5.6 Fair Representations under Multiple Notions

Finally, we demonstrate how L-MIFR could control multiple fairness constraints simultaneously, thereby finding representations that are reasonably fair when there are multiple notions in question. We consider the *Adult* dataset, and describe the *demographic parity*, *equalized odds* and *equalized opportunity* notions of fairness in terms of mutual information, which we denote as $I_{DP} := I_q(\mathbf{z}; \mathbf{u})$, I_{EO}, I_{EOpp} respectively (see details in Appendix C about how I_{EO} and I_{EOpp} are derived).

	$I_q(\mathbf{x};\mathbf{z}\|\mathbf{u})$	C_1	I_{DP}	I_{EO}	I_{EOpp}
MIFR	9.34	**9.39**	0.09	0.10	0.07
L-MIFR	**9.94**	9.95	**0.08**	**0.09**	**0.04**

Table 3: Learning one representation for multiple notions of fairness on *Adult*. L-MIFR learns representations that are better than MIFR on all the measurements instead of C_1. Here $\epsilon_1 = 10$ for C_1 and $\epsilon = 0.1$ for other constraints.

For L-MIFR, we set $\epsilon_1 = 10$ and other ϵ values to 0.1. For MIFR, we consider a more efficient approach than random grid search. We start by setting every $\lambda = 0.1$; then we multiply the λ value for a particular constraint by 2 until the constraint is satisfied by MIFR; we finish when all the constraints are satisfied[7]. We find that this requires us to update the λ of I_{DP}, I_{EO} and I_{EOpp} four times each (so corresponding $\lambda = 1.6$); this cost 12x the computational resources needed by L-MIFR.

We compare the representations learned by L-MIFR and MIFR in Figure 3. L-MIFR outperforms MIFR in terms of $I_q(\mathbf{x};\mathbf{z}|\mathbf{u})$, I_{DP}, I_{EO} and I_{EOpp}, while only being slightly worse in terms of C_1. Since $\epsilon_1 = 10$, the L-MIFR solution is still feasible. This demonstrates that even with a hand-designed method for tuning λ, MIFR is still vastly inferior to L-MIFR in terms of computational cost and representation expressiveness.

6. DISCUSSION

In this paper, we introduced an objective for learning controllable fair representations based on mutual information. This interpretation allows us to unify and explain existing work. In particular, we have shown that a range of existing approaches optimize an approximation to the Lagrangian dual of our objective with *fixed* multipliers, fixing the trade-off between fairness and expressiveness. We proposed a dual optimization method that allows us to achieve higher expressiveness while satisfying the user-specified limit on unfairness.

Our method is also related to other works on fairness and information theory. Komiyama et al. (2018) solve least square regression under multiple fairness constraints. Calmon et al. (2017) transform the dataset to prevent discrimination on specific classification tasks. Our work is most related to Zhao et al. (2018), which discussed information-theoretic constraints in the context of learning latent variable generative models.

In future work, we are interested in formally and empirically extending this framework (and the corresponding dual optimization method) to other notions of fairness. It is also valuable to investigate alternative approaches to training the adversary (Gulrajani et al., 2017), the usage of more flexible $p(\mathbf{z})$ (Rezende and Mohamed, 2015), and alternative solutions to bounding $I_q(\mathbf{z},\mathbf{u})$.

Bibliography

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, December 2016.

Alexander A Alemi, Ben Poole, Ian Fischer, Joshua V Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. *arXiv preprint arXiv:1711.00464*, November 2017.

T Calders, F Kamiran, and M Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, December 2009. doi: 10.1109/ICDMW.2009.83.

7. This allows MIFR to approach the feasible set from outside, so the solution it finds will generally have high expressiveness.

Flavio P Calmon, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized data Pre-Processing for discrimination prevention. *arXiv preprint arXiv:1704.03354*, April 2017.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational lossy autoencoder. *arXiv preprint arXiv:1611.02731*, November 2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, June 2017. ISSN 2167-647X, 2167-6461. doi: 10.1089/big.2016.0047.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, May 2016.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, November 2015.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

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

Diederik P Kingma and Max Welling. Auto-Encoding variational bayes. *arXiv preprint arXiv:1312.6114v10*, December 2013.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, June 2016.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent Trade-Offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, September 2016.

Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2742–2751, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, November 2015.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep Latent-Variable models. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6446–6456. Curran Associates, Inc., 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, February 2018.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, May 2015.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, January 2016.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Learning fair classifiers. *arXiv preprint arXiv:1507. 05259*, 2015.

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

Shengjia Zhao, Jiaming Song, and Stefano Ermon. InfoVAE: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, June 2017a.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Towards deeper understanding of variational autoencoding models. *arXiv preprint arXiv:1702.08658*, February 2017b.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. A lagrangian perspective to latent variable generative models. *Conference on Uncertainty in Artificial Intelligence*, 2018.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, May 2015.

Appendix A. Proofs

A.1 Proof of Lemma 1

Proof

$$\begin{aligned}
I_q(\mathbf{x};\mathbf{z}|\mathbf{u}) &= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{x},\mathbf{z}|\mathbf{u}) - \log q(\mathbf{x}|\mathbf{u}) - \log q_\phi(\mathbf{z}|\mathbf{u})] \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{x},\mathbf{z}|\mathbf{u}) - \log q_\phi(\mathbf{z}|\mathbf{u})] + \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[-\log q(\mathbf{x}|\mathbf{u})] \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{x}|\mathbf{z},\mathbf{u})] + H_q(\mathbf{x}|\mathbf{u}) \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{x}|\mathbf{z},\mathbf{u}) + \log p(\mathbf{x}|\mathbf{z},\mathbf{u}) - \log p(\mathbf{x}|\mathbf{z},\mathbf{u})] + H_q(\mathbf{x}|\mathbf{u}) \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p(\mathbf{x}|\mathbf{z},\mathbf{u})] + H_q(\mathbf{x}|\mathbf{u}) + \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{x}|\mathbf{z},\mathbf{u}) \| p(\mathbf{x}|\mathbf{z},\mathbf{u})) \\
&\geq \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p(\mathbf{x}|\mathbf{z},\mathbf{u})] + H_q(\mathbf{x}|\mathbf{u})
\end{aligned}$$

where the last inequality holds because KL divergence is non-negative. ■

A.2 Proof of Lemma 2

Proof

$$\begin{aligned}
I_q(\mathbf{z};\mathbf{u}) &\leq I_q(\mathbf{z};\mathbf{x},\mathbf{u}) \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) - \log q_\phi(\mathbf{z})] \\
&= \mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) - \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}) + \log p(\mathbf{z})] \\
&= \mathbb{E}_{q(\mathbf{x},\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) \| p(\mathbf{z})) - D_{\mathrm{KL}}(q_\phi(\mathbf{z}) \| p(\mathbf{z}))
\end{aligned}$$

■

A.3 Proof of Lemma 3

Proof

$$\begin{aligned}
I_q(\mathbf{z};\mathbf{u}) &= \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{u}|\mathbf{z}) - \log q(\mathbf{u})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u})}[\log q_\phi(\mathbf{u}|\mathbf{z}) - \log p(\mathbf{u}) - \log q(\mathbf{u}) + \log p(\mathbf{u})] \\
&= \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) - D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u})) \\
&\leq \mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u}))
\end{aligned}$$

Again, the last inequality holds because KL divergence is non-negative. ■

A.4 Proof of Theorem 5

Proof Let us first verify that this problem is convex.

- Primal: $-\mathbb{E}_{q_\phi(\mathbf{x},\mathbf{z},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})]$ is affine in $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})$, convex in $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})$ due to the concavity of log, and independent of $p_\theta(\mathbf{z})$.
- First condition: $\mathbb{E}_{q(\mathbf{u})} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u}) \| p_\theta(\mathbf{z}))$ is convex in $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})$ and $p_\theta(\mathbf{z})$ (because of convexity of KL-divergence), and independent of $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{u})$.

- Second condition: since $\mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) - D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u})) = I_q(\mathbf{z}; \mathbf{u})$ and

$$I_q(\mathbf{z}; \mathbf{u}) = D_{\mathrm{KL}}(q_\phi(\mathbf{z}, \mathbf{u}) \| q(\mathbf{u}) q_\phi(\mathbf{z})) \tag{15}$$
$$= D_{\mathrm{KL}}(\sum_{\mathbf{x}} q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u}) \| q(\mathbf{u}) \sum_{\mathbf{x}, \mathbf{u}} q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u})) \tag{16}$$

Let $q = \beta q_1 + (1 - \beta) q_2$, $\forall \beta \in [0, 1], q_1, q_2$. We have

$$\begin{aligned} I_q(\mathbf{z}; \mathbf{u}) &= D_{\mathrm{KL}}(\sum_{\mathbf{x}} q(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u}) \| q(\mathbf{u}) \sum_{\mathbf{x}, \mathbf{u}} q(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u})) \\ &\geq \beta D_{\mathrm{KL}}(\sum_{\mathbf{x}} q_1(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u}) \| q(\mathbf{u}) \sum_{\mathbf{x}, \mathbf{u}} q_1(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u})) \\ &\quad + (1 - \beta) D_{\mathrm{KL}}(\sum_{\mathbf{x}} q_2(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u}) \| q(\mathbf{u}) \sum_{\mathbf{x}, \mathbf{u}} q_2(\mathbf{z}|\mathbf{x}, \mathbf{u}) q(\mathbf{x}, \mathbf{u})) \\ &= \beta I_{q_1}(\mathbf{z}; \mathbf{u}) + (1 - \beta) I_{q_2}(\mathbf{z}; \mathbf{u}) \end{aligned}$$

where we use the convexity of KL divergence in the inequality. Since $D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u}))$ is independent of $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$, both $I_q(\mathbf{z}; \mathbf{u})$ and $\mathbb{E}_{q_\phi(\mathbf{z})} D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u}))$ are convex in $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$.

Then we show that the problem has a feasible solution by construction. In fact, we can simply let $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) = p_\theta(\mathbf{z})$ be some fixed distribution over $\mathbf{z}$, and $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u}) = q_\phi(\mathbf{x}|\mathbf{z}, \mathbf{u})$ for all $\mathbf{x}, \mathbf{u}$. In this case, $\mathbf{z}$ and $\mathbf{u}$ are independent, so $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}) \| p_\theta(\mathbf{z})) = 0 < \epsilon_1$, $D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}) \| p(\mathbf{u})) = 0 < \epsilon_2$. This corresponds to the case where $\mathbf{z}$ is simply random noise that does not capture anything in $\mathbf{u}$.

Hence, Slater's condition holds, which is a sufficient condition for strong duality. ∎

Appendix B. Experimental Setup Details

We consider the following setup for our experiments.

- For MIFR, we modify the weight for reconstruction error $\alpha = 1$, as well as $\lambda_1 \in \{0.0, 0.1, 0.2, 1.0, 2.0\}$ and $\lambda_2 \in \{0.1, 0.2, 1.0, 2.0, 5.0\}$ for the constraints, which creates a total of $5^2 = 25$ configurations; λ_1 values smaller since high values of λ_1 prefers solutions with low $I_q(\mathbf{x}; \mathbf{z}|\mathbf{u})$.
- For L-MIFR, we modify ϵ_1 and ϵ_2 according to the estimated values for each dataset. This allows us to claim results that holds for a certain hyperparameter in general (even as other hyperparameter change).
- We use the Adam optimizer with initial learning rate $1e - 3$ and $\beta_1 = 0.5$ where the learning rate is multiplied by 0.98 every 1000 optimization iterations, following common settings for adversarial training (Gulrajani et al., 2017).
- For L-MIFR, we initialize the λ_i parameters to 1.0, and allow for a range of $(0.01, 100)$.
- Unless otherwise specified, we update $p_\psi(\mathbf{u}|\mathbf{z})$ 10 times per update of $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})$ and $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u})$.
- For *Adult* and *Health* we optimize for 2000 epochs; for *German* we optimize for 10000 epochs (since there are only 1000 low dimensional data points).
- For both cases, we consider $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u}), p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{u}), p_\psi(\mathbf{u}|\mathbf{z})$ as a two layer neural networks with a hidden layer of 50 neurons with softplus activations, and use $\mathbf{z}$ of dimension 10 for *German* and *Adult*, and 30 for *Health*. For the joint of two variables (i.e. $(\mathbf{x}, \mathbf{u})$) we simply concatenate them at the input layer. We find that our conclusions are insensitive to a reasonable change in architectures (e.g. reduce number of neurons to 50 and $\mathbf{z}$ to 25 dimensions).

We show a comparison of Δ_{DP}, Δ_{EO}, Δ_{EOpp} between L-MIFR and LAFTR (Madras et al., 2018) on the Adult dataset in Table 4, where L-MIFR is trained with the procedure in Section 5.6. While LAFTR achieves better fairness on each notion if it is specifically trained for that notion, it often achieves worse performance on other notions of fairness. We note that L-MIFR uses a logistic regression classifier, whereas LAFTR uses a one layer MLP. Moreover, these measurements are also task-specific as opposed to mutual information criterions.

	Δ_{DP}	Δ_{EO}	Δ_{EOpp}
L-MIFR	0.057	0.123	0.026
LAFTR-DP	0.029	0.244	0.027
LAFTR-EO	0.125	0.074	0.037
LAFTR-EOpp	0.098	0.154	0.022

Table 4: Comparison between L-MIFR and LAFTR on Δ_{DP}, Δ_{EO}, Δ_{EOpp} metrics from (Madras et al., 2018). While LAFTR achieves better fairness on individual notions if it is trained for that notion, it often trades that with other notions of fairness.

Appendix C. Extension to Equalized Odds and Equalized Opportunity

If we are also provided labels y for a particular task, in the form of $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{u}_i, y_i)\}_{i=1}^M$, we can also use the representations to predict y, which leads to a third condition:

3. **Classification** $\mathbf{z}$ can be used to classify y with high accuracy.

We can either add this condition to the primal objective in Equation 1, or add an additional constraint that we wish to have accuracy that is no less than a certain threshold.

With access to binary labels, we can also consider information-theoretic approaches to *equalized odds* and *equalized opportunity* (Hardt et al., 2016). Recall that *equalized odds* requires that the predictor and sensitive attribute are independent conditioned on the label, whereas *equalized opportunity* requires that the predictor and sensitive attribute are independent conditioned on the label being positive. In the case of learning representations for downstream tasks, our notions should consider any classifier over $\mathbf{z}$.

For *equalized odds*, we require that z and u have low mutual information conditioned on the label, which is $I_q(\mathbf{z}, \mathbf{u}|y)$. For *equalized opportunity*, we require that z and u have low mutual information conditioned on the label $y = 1$, which is $I_q(\mathbf{z}, \mathbf{u})|_{y=1}$.

We can still apply the upper bounds similar to the case in C_2. For *equalized opportunity* we have

$$\begin{aligned} I_q(\mathbf{z}; \mathbf{u})|_{y=1} &\leq \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u},y|y=1)}[D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}, y) \| p(\mathbf{u}))] - D_{\mathrm{KL}}(q(\mathbf{u}) \| p(\mathbf{u})) := I_{EO} \\ &\leq \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u},y|y=1)}[D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}, y) \| p(\mathbf{u}))] \end{aligned}$$

For *equalized odds* we have

$$\begin{aligned} I_q(\mathbf{z}; \mathbf{u}|y) &= q(1) I_q(\mathbf{z}; \mathbf{u})|_{y=1} + q(0) I_q(\mathbf{z}; \mathbf{u})|_{y=0} := I_{EOpp} \\ &\leq q(1)\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u},y|y=1)}[D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}, y) \| p(\mathbf{u}))] + q(0)\mathbb{E}_{q_\phi(\mathbf{z},\mathbf{u},y|y=0)}[D_{\mathrm{KL}}(q_\phi(\mathbf{u}|\mathbf{z}, y) \| p(\mathbf{u}))] \end{aligned}$$

which can be implemented by using a separate classifier for each y or using y as input. If y is an input to the classifier, our mutual information formulation of *equalized odds* does not have to be restricted to the case where y is binary.