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. □

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).

C Comparison with LAFTR

Our work have several notable differences from prior methods (such as LAFTR (Madras et al., 2018)) that make it hard to compare them directly. First, we do not assume access to the prediction task while learning the representation, thus our method does not directly include the "classification error" objective. Second, our method is able to deal with any type of sensitive attributes, as opposed to binary ones.

Nevertheless, we compare the performance of MIFR and LAFTR (Madras et al.) with the demographic parity notion of fairness (measured by $Delta_{DP}$, lower is better). To make a fair comparison, we add a classification

error to MIFR during training. MIFR achieves an accuracy of 0.829 and Δ_{DP} of 0.037, whereas LAFTR achieves an accuracy of 0.821 and Δ_{DP} of 0.029. This shows that MIFR and LAFTR are comparable in terms of the accuracy / fairness trade-off. MIFR is still useful for sensitive attributes that are not binary, such as Health, which LAFTR cannot handle.

We further 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.

D 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 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.