Controllable Guarantees for Fair Outcomes via Contrastive Information Estimation

Umang Gupta[1] , **Aaron Ferber**[2] , **Bistra Dilkina**[2] , **Greg Ver Steeg**[1]
[1] USC Information Sciences Institute
[2] University of Southern California
{umanggup, aferber, dilkina}@usc.edu, {gregv}@isi.edu

Abstract

Controlling bias in training datasets is vital for ensuring equal treatment, or parity, between different groups in downstream applications. A naive solution is to transform the data so that it is statistically independent of group membership, but this may throw away too much information when a reasonable compromise between fairness and accuracy is desired. Another common approach is to limit the ability of a particular adversary who seeks to maximize parity. Unfortunately, representations produced by adversarial approaches may still retain biases as their efficacy is tied to the complexity of the adversary used during training. To this end, we theoretically establish that by limiting the mutual information between representations and protected attributes, we can assuredly control the parity of any downstream classifier. We demonstrate an effective method for controlling parity through mutual information based on contrastive information estimators and show that it outperforms other existing approaches. We test our approach on *UCI Adult* and *Heritage Health* datasets and show that our approach provides more informative representations across a range of desired parity thresholds while providing strong theoretical guarantees on the parity of any downstream algorithm.

1 Introduction

Learning algorithms often exploit and exaggerate biases present in the training dataset. To tackle this problem, a set of approaches propose learning representations or preprocessing the data in a way that removes information about the protected attributes ensuring any downstream algorithm cannot use the sensitive information [Song *et al.*, 2019; McNamara *et al.*, 2017]. Ideally, users of this transformed data can focus on maximizing performance for their tasks using any methods available without the risk of producing biased or unfair outcomes [Cisse and Koyejo, 2019]. The strongest requirement for fair representations is to be statistically independent of sensitive attributes, but this may lead to large drops in predictive performance as sensitive attributes are often correlated with the target. Therefore, it is desirable to produce representations that can trade-off some measure of fairness with the utility [Menon and Williamson, 2018; Dutta *et al.*, 2019].

Representative Methods	Adversarial Guarantee	Controllable Parity
Song *et al.* [2019]	Weak[1]	Heuristic
Moyer *et al.* [2018]	**Strong**	No[2]
Madras *et al.* [2018]	None	Heuristic
Roy and Boddeti [2019]	None	No
Ours	**Strong**	**Provable**

Table 1: Fair representation learning methods

Many recent approaches for learning fair representations leverage adversarial learning to remove unwanted biases from the data by using an adversary during training that tries to reconstruct sensitive attributes from the representation [Jaiswal *et al.*, 2020; Roy and Boddeti, 2019; Madras *et al.*, 2018]. While adversarial methods have been shown to be useful in learning fair representations, they are often limited by the adversary's model capacity [Xu *et al.*, 2020]. A predictive model more powerful than the adversary used in training may reveal hidden biases that are present in the representations. Hence, a model trained with adversarial learning has no guarantee to control fairness against an arbitrary adversary. Other methods for learning fair representations focus on the stricter constraint of inducing statistical independence [Moyer *et al.*, 2018; Louizos *et al.*, 2016] and not on trading-off between fairness and informativeness. We summarize some of these approaches and their properties in Table 1.

Our main contributions are - a) theoretically show that mutual information between representation and sensitive-attributes bounds the parity of any decision algorithm, and b) propose practical ways to limit mutual information leveraging contrastive information estimators that can efficiently trade-off predictability and accuracy.

[1] Song *et al.* [2019] minimize two different bounds on $I(\mathbf{z} : \mathbf{c})$ — one is a very loose upper bound and another uses adversarial learning. So the adversarial guarantee is unclear or at best weaker.

[2] Moyer *et al.* [2018] designed their method for enforcing independence, however we consider a modification of their method for controlling parity, based on our Theorem 2.

2 Mutual Information Bounds Parity

We consider a dataset of triplets $\mathcal{D} = \{x_i, y_i, c_i\}_{i=1}^N$, where x_i, y_i, c_i are iid samples from data distribution $p(\mathbf{x}, \mathbf{y}, \mathbf{c})$. $\mathbf{c}$ are the sensitive or protected attributes, $\mathbf{y}$ is the label, $\mathbf{x}$ are features of the sample which may include sensitive attributes and $\hat{\mathbf{y}}$ denotes predicted label, according to some algorithm. We may also interpret $\hat{\mathbf{y}}$ as the outcome of some decision procedure. We use bold letters to denote the random variable, and the regular font represents corresponding samples. In this work, we consider stochastic representations of data i.e. $z(x) \sim q(\mathbf{z} \mid \mathbf{x} = x)$. We want to learn d-dimensional representations z of input x, such that any classifier trained on only z is guaranteed to be fair, i.e., it has statistical parity within some δ'. In this work, we focus on statistical parity, a popular measure of group fairness, and it is defined as:

Definition 1. ***Statistical Parity:*** *[Dwork* et al.*, 2012] It is the absolute difference between the selection rates of two groups. Mathematically,*

$$\Delta_{DP}(\mathcal{A}, \mathbf{c}) = |P(\hat{\mathbf{y}} = 1 \mid \mathbf{c} = 1) - P(\hat{\mathbf{y}} = 1 \mid \mathbf{c} = 0)|$$

where $\hat{\mathbf{y}}$ denotes decisions produced by some decision algorithm $\mathcal{A}$. When there are more than two groups, we define statistical parity to be the maximum parity between any two groups (as implemented in Bird et al. *[2020]).*

Mutual information between representations and protected attributes, denoted as $I(\mathbf{z}:\mathbf{c})$, can be used to limit the statistical parity via the following result.

Theorem 2. *For some $z, c \sim p(\mathbf{z}, \mathbf{c})$, $z \in \mathbb{R}^d$, $c \in \{0, 1\}$, and any decision algorithm $\mathcal{A}$ that acts on z, we have*

$$I(\mathbf{z}:\mathbf{c}) \geq g\left(\pi, \Delta_{DP}\left(\mathcal{A}, \mathbf{c}\right)\right)$$

where $\pi = P(\mathbf{c} = 1)$ and g is monotonically increasing.

We omit the proof of the theorem and exact form of g due to space constraints. We only require the following information about g for our arguments—g is monotonically increasing in Δ_{DP}, and π is constant for a specific dataset. We see from Thm. 2 that $I(\mathbf{z}:\mathbf{c})$ bounds the parity of any downstream decision algorithm. Since g is a monotonically increasing function, any reduction in $I(\mathbf{z}:\mathbf{c})$ will decrease Δ_{DP} too.

We emphasize that $I(\mathbf{z}:\mathbf{c})$ is often used as a proxy objective to control statistical parity [Edwards and Storkey, 2016; Song *et al.*, 2019; Moyer *et al.*, 2018]. It is often justified via the data processing inequality and the intuition that both statistical parity and mutual information are measures of dependence. However, due to data processing inequality, we can only guarantee that if we limit information about $\mathbf{c}$ in $\mathbf{z}$, then no subsequent operations on $\mathbf{z}$ can increase information about $\mathbf{c}$, i.e., $I(\hat{\mathbf{y}}:\mathbf{c}) \leq I(\mathbf{z}:\mathbf{c})$, but the effect on statistical parity is unclear. Our result (Thm. 2) demonstrates that limiting mutual information can monotonically limit statistical parity, which had not been theoretically demonstrated until now.

3 Practical Objectives for Controlling Parity

Equipped with an algorithm agnostic upper bound to parity, we now discuss practical objectives for learning fair representations. Along with limiting parity, we also want the latent representation to be highly predictive (informative) about the label, which is often realized by maximizing mutual information between $\mathbf{y}$ and $\mathbf{z}$, i.e., $I(\mathbf{y}:\mathbf{z})$ implicitly [Edwards and Storkey, 2016; Madras *et al.*, 2018; Jaiswal *et al.*, 2020] or explicitly [Moyer *et al.*, 2018].

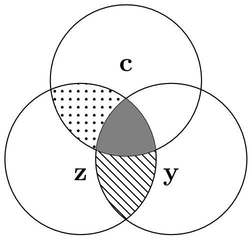

Figure 1: Venn diagram to show interference between $I(\mathbf{z}:\mathbf{c})$ and $I(\mathbf{z}:\mathbf{y})$. The dotted and dashed part are minimized or maximized, respectively, but the gray region is both minimized and maximized. See Sec. 3.1 for details.

$$\begin{aligned} \mathcal{O}_1 : \max_q I(\mathbf{y}:\mathbf{z}) \;\; \text{s.t.} \;\; I(\mathbf{z}:\mathbf{c}) \leq \delta \\ \text{or, } \max_q I(\mathbf{y}:\mathbf{z}) - \beta I(\mathbf{z}:\mathbf{c}) \end{aligned} \tag{1}$$

where, $\mathbf{z}, \mathbf{x} \sim q(\mathbf{z} \mid \mathbf{x})p(\mathbf{x})$ and $\beta > 0$ is a hyperparameter.

3.1 Interference between $I(\mathbf{y}:\mathbf{z})$ and $I(\mathbf{z}:\mathbf{c})$

While $I(\mathbf{y}:\mathbf{z})$ has been commonly proposed as a criterion to enforce the desiderata of representations being informative about labels, when the data is biased, i.e., $I(\mathbf{y}:\mathbf{c}) > 0$, maximizing $I(\mathbf{y}:\mathbf{z})$ is in direct contradiction with minimizing $I(\mathbf{z}:\mathbf{c})$. To illustrate this point, we refer to the information Venn diagram in Fig. 1. The goal of fair representation learning is to move the circle representing information about the representation, $\mathbf{z}$, to have high overlap with $\mathbf{y}$ and low overlap with $\mathbf{c}$. However, there is a conflict in the gray region where we cannot increase overlap with $\mathbf{y}$ without also increasing overlap with $\mathbf{c}$. We observe experimentally that this conflict hurts the model performance and makes it hard to achieve lower parity values at a reasonable accuracy. This conflict can be avoided. Since fair learning aims to capture information about $\mathbf{y}$ that is *not* related to the protected attribute $\mathbf{c}$, we want to maximize the overlap between $\mathbf{z}$ and the region of $\mathbf{y}$ that excludes $\mathbf{c}$. This quantity is precisely the conditional mutual information, $I(\mathbf{y}:\mathbf{z} \mid \mathbf{c})$, which we propose to maximize instead of $I(\mathbf{y}:\mathbf{z})$. This leads us to the following objective:

$$\begin{aligned} \mathcal{O}_2 : \max_q I(\mathbf{y}:\mathbf{z} \mid \mathbf{c}) \;\; \text{s.t.} \;\; I(\mathbf{z}:\mathbf{c}) \leq \delta \\ \text{or, } \max_q I(\mathbf{y}:\mathbf{z} \mid \mathbf{c}) - \beta I(\mathbf{z}:\mathbf{c}) \end{aligned} \tag{2}$$

where, $\mathbf{z}, \mathbf{x} \sim q(\mathbf{z} \mid \mathbf{x})p(\mathbf{x})$ and $\beta > 0$ is a hyperparameter. Eq. 2 defines our approach, but the information-theoretic terms are difficult to estimate directly. Next, we derive practical variational bounds for the terms appearing in Eqs. 1 & 2.

3.2 Lower bounds for $I(\mathbf{y}:\mathbf{z})$ and $I(\mathbf{y}:\mathbf{z} \mid \mathbf{c})$

$$\begin{aligned} I(\mathbf{y}:\mathbf{z}) &= H(\mathbf{y}) - H(\mathbf{y} \mid \mathbf{z}) \\ &= H(\mathbf{y}) + \mathbb{E}_{\mathbf{y},\mathbf{z}} \log r(\mathbf{y} \mid \mathbf{z}) + \text{KL}\left(p(\mathbf{y} \mid \mathbf{z}) \parallel r(\mathbf{y} \mid \mathbf{z})\right) \\ &\geq H(\mathbf{y}) + \max_r \mathbb{E}_{\mathbf{y},\mathbf{z}} \log r(\mathbf{y} \mid \mathbf{z}) \end{aligned} \tag{3}$$

and similarly,

$$I(\mathbf{y}:\mathbf{z} \mid \mathbf{c}) \geq H(\mathbf{y} \mid \mathbf{c}) + \max_r \mathbb{E}_{\mathbf{y},\mathbf{z},\mathbf{c}} \log r(\mathbf{y} \mid \mathbf{z}, \mathbf{c}) \tag{4}$$

$H(\mathbf{y})$ and $H(\mathbf{y} \mid \mathbf{c})$ are properties of data and, therefore, constant from the optimization perspective. When $\mathbf{y}$ is a one-dimensional variable denoting the target class, this is equivalent to minimizing cross-entropy. To this end, we will parametrize the variational distribution r using a neural network with parameters ψ, but other models can also be used.

3.3 Upper bound for $I(\mathbf{z} : \mathbf{c})$

Our technique for upper-bounding $I(\mathbf{z}:\mathbf{c})$ is similar to Moyer *et al.* [2018] and makes use of the following observation:

$$I(\mathbf{z}:\mathbf{c}) = I(\mathbf{z}:\mathbf{c} \mid \mathbf{x}) + I(\mathbf{z}:\mathbf{x}) - I(\mathbf{z}:\mathbf{x} \mid \mathbf{c}) \qquad (5)$$

$I(\mathbf{z}:\mathbf{c} \mid \mathbf{x}) = 0$, as z is a function of x and some independent noise. As a result, we have $I(\mathbf{z}:\mathbf{c}) = I(\mathbf{z}:\mathbf{x}) - I(\mathbf{x}:\mathbf{z} \mid \mathbf{c})$. The first term is the information bottleneck term [Alemi *et al.*, 2017] and limits the information about $\mathbf{x}$ in $\mathbf{z}$, and we will bound it by specifying a prior over $\mathbf{z}$. The second term tries to preserve information about $\mathbf{x}$ but not in $\mathbf{c}$ and we will lower bound it via contrastive estimation .

Upper bound for $I(\mathbf{z}:\mathbf{x})$ by specifying a prior: In order to upper-bound $I(\mathbf{z}:\mathbf{x})$, we use the following result:

$$\begin{aligned} I(\mathbf{z}:\mathbf{x}) &= \mathbb{E}_{\mathbf{x}}\mathrm{KL}\left(q(\mathbf{z} \mid \mathbf{x};\phi) \parallel p(\mathbf{z})\right) - \mathrm{KL}\left(p(\mathbf{z}) \parallel q(\mathbf{z})\right) \\ &\leq \mathbb{E}_{\mathbf{x}}\mathrm{KL}\left(q(\mathbf{z} \mid \mathbf{x};\phi) \parallel p(\mathbf{z})\right) \qquad (6) \end{aligned}$$

where $p(\mathbf{z})$ is any distribution. This is similar to the rate term in a VAE or information bottleneck approach [Alemi *et al.*, 2017; Higgins *et al.*, 2017]. Motivated by this similitude, we let $p(\mathbf{z})$ be standard normal distribution and $q(\mathbf{z} \mid \mathbf{x};\phi)$ be a diagonal gaussian distribution whose mean and variance are parametrized as $\mu(x) = f_\mu(x;\phi), \Sigma(x) = f_\Sigma(x;\phi)$.

Lower bound for $I(\mathbf{x}:\mathbf{z} \mid \mathbf{c})$ via constrative estimation: We propose to lower bound $I(\mathbf{x}:\mathbf{z} \mid \mathbf{c})$ via contrastive mutual information estimation by using the following proposition.

Proposition 3. *[Poole* et al.*, 2019] For any $u, v, w \sim p(\mathbf{u}, \mathbf{v}, \mathbf{w})$, $\tilde{u} \sim p(\mathbf{u} \mid \mathbf{w})$, and function f, we have*

$$I(\mathbf{u}:\mathbf{v} \mid \mathbf{w}) \geq \mathbb{E}_{\mathbf{u},\mathbf{v},\mathbf{w}} \log \frac{e^{f(u,v,w)}}{\frac{1}{M}\sum_{j=1}^{M} e^{f(\tilde{u}_j,v,w)}} \qquad (7)$$

where, $u, \tilde{u}_j \in \mathcal{U}$, $v \in \mathcal{V}$, $w \in \mathcal{W}$, $f : \mathcal{U} \times \mathcal{V} \times \mathcal{W} \to \mathbb{R}$, and M is the number of samples from $p(\mathbf{u} \mid \mathbf{w})$.

As a direct application of this result, we can lower-bound and maximize $I(\mathbf{x}:\mathbf{z} \mid \mathbf{c})$. However, there is a caveat that we need to sample from $p(\mathbf{z} \mid \mathbf{c})$. Sampling from this conditional distribution in general case can be hard but for our problem, it can be easily accomplished. Often for fairness applications, $\mathbf{c}$ is a discrete random variable with low cardinality. Infact it is often a binary random variable. Therefore, $\{z_j : (z_j, c_j = i)\}$ can be considered as samples from $p(\mathbf{z} \mid \mathbf{c} = i)$. In our experiments, we parametrize $f(z, x, c)$ as a bilinear function (similar to Oord *et al.* [2018]) and $f(z, x, c) = z^T \mathbf{W}^T e(x;\theta')$, where $\mathbf{W}, \theta'$ are learnable parameters.

We use results from Eqs. 4, 6, 7 to tractably compute and maximize the proposed objective in Eq. 2. We call our objective **F**air **C**ontrastive **R**epresentation **L**earner (FCRL).

4 Experiments

Datasets: We validate our approach on *UCI Adult* [Dua and Graff, 2017] and *Heritage Health*[3] Dataset. *UCI Adult* is 1994 census data with train and test set size of 30K and 15K, respectively. The target task is to predict whether the income exceeds $50K, and the protected attribute is considered gender (which is binary in this case). We use the same preprocessing as Moyer *et al.* [2018]. *Heritage Health* dataset is data of around 51K patients (40K in the train set and 11K in the test set), and the task is to predict the Charleson Index, which is an indicator of 10-year survival of a patient. We consider age as the protected attribute, which has 9 possible values. We use the same preprocessing as Song *et al.* [2019].

Evaluation Procedure: A fair representation learning algorithm aims to produce representations such that any downstream decision algorithm that uses these representations will produce fairer results. Therefore similar to Madras *et al.* [2018], we train the representation learning algorithm on training data and evaluate the representations by training classifiers for downstream prediction tasks. Since our purpose is to assess the representations, we report average accuracy (as an indicator of most likely performance) and maximum parity (as an indicator of worst-case bias) computed over 5 runs of the decision algorithm with random seeds. Unlike Madras *et al.* [2018], we also allow for preprocessing to be done on representations. Preprocessing steps like min-max or standard scaling are common and often precede training of classifiers in a regular machine learning pipeline.

Baselines and Architecture: We compare with a number of recent approaches, including information-theoretic and adversarial methods from the recent literature. Among the information-theoretic methods, we compare with MIFR [Song *et al.*, 2019] which generalizes several previous fair representation learning approaches [Louizos *et al.*, 2016; Edwards and Storkey, 2016; Madras *et al.*, 2018; Zemel *et al.*, 2013] and CVIB [Moyer *et al.*, 2018]. We also compare with recent state-of-the-art adversarial methods of Jaiswal *et al.* [2020] (Adversarial Forgetting), Roy and Boddeti [2019] (MaxEnt-ARL) and Madras *et al.* [2018] (LAFTR). LAFTR is only applicable when $\mathbf{c}$ is a binary variable. As a baseline, we also train a one hidden layer MLP predictor directly on the data without regards to fairness (*Unfair MLP*). For a fair comparison, we set $d = 8$ for all the methods and use model components like encoder, decoder, etc. of the same complexity. We use a 1-hidden layer MLP with ReLU non-linearity and 50 neurons in the hidden layer as our choice of decision algorithm, and representations are preprocessed by standard scaling.

Improved Accuracy vs. Parity Trade-offs: For different fair representation learners, we compare accuracy versus parity achieved for the above-specified classifier. We visualize the trade-offs between fairness and task performance by plotting parity vs. accuracy curves for each representation learning algorithm by varying each method's inherent hyperparameters over the range specified in the original works to get different points on this curve in Fig. 2. The goal is to push the frontier of achievable trade-offs as far to the bottom-right as possible, i.e., to achieve the best possible accuracy while maintaining a low parity. From a visual inspection of Fig. 2, we can see that our approach preserves more information about label $\mathbf{y}$, across a range of fairness thresholds for both the datasets.

For the Adult dataset, MIFR [Song *et al.*, 2019] is compet-

[3] https://www.kaggle.com/c/hhp

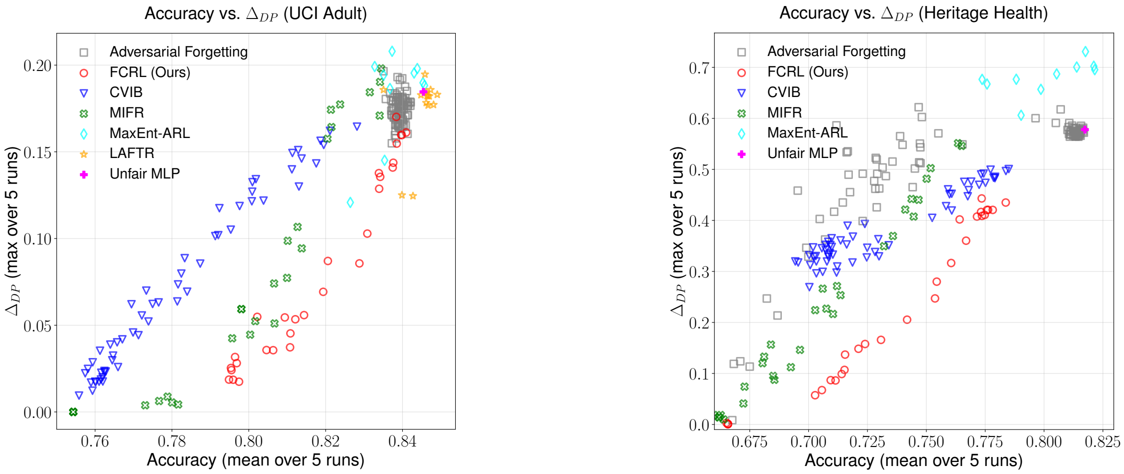

Figure 2: Parity vs. Accuracy trade-off for *UCI Adult* and *Heritage Health* dataset using a 1 hidden layer MLP. Lower Δ_{DP} is better and higher accuracy is better. To get different trade-off points, we use representations generated by varying each method's inherent loss hyperparameters.

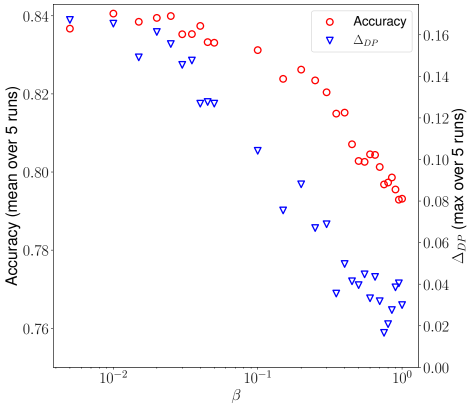

Figure 3: Parity and Accuracy variation with β. Our method can explore feasible regions of parity and accuracy by varying only a single parameter β

itive with our approach near low parity; however, it fails to achieve higher accuracy at higher demographic parity. This is because MIFR proposed to use $I(\mathbf{x}:\mathbf{z})$ as an upper bound to $I(\mathbf{z}:\mathbf{c})$, which is very loose, and this penalizes information about x as well, which is not desirable. CVIB [Moyer *et al.*, 2018] is able to consistently trade-off accuracy while using a reconstruction based bound. But it maximizes $I(\mathbf{y}:\mathbf{z})$ which conflicts with desired minimization of $I(\mathbf{z}:\mathbf{c})$ (see Sec. 3.1).

Controlling Parity: Our approach has a single intuitive hyperparameter β, which can be used to control $I(\mathbf{z}:\mathbf{c})$ directly and, therefore, via Thm. 2, to monotonically control parity (see Fig. 3).

5 Related Work

Fair classification methods are often categorized based on which stage of the machine learning pipeline they target. Our approach targets the pre-processing stage. Pre-processing methods are useful when the onus of fairness is on a third party or the data controller, and the end-user may be oblivious to fairness constraints [McNamara *et al.*, 2017; Cisse and Koyejo, 2019]. Pre-processing methods must ensure fairness with respect to any downstream classification algorithm. Many pre-processing methods have discussed the desiderata of ensuring strong guarantees on fairness so that any downstream classifier may be used freely [McNamara *et al.*, 2017; Song *et al.*, 2019; Edwards and Storkey, 2016; Madras *et al.*, 2018]. However, their operationalization often leads to an approach that may not ensure guarantees (due to limits of adversarial methods, for instance). Other works have explored information-theoretic objectives for learning invariant representations [Moyer *et al.*, 2018] and fair representations [Song *et al.*, 2019]. Dutta *et al.* [2019] use tools from information theory to analyze the trade-off between fairness and accuracy.

Contrastive learning and its variants have shown promising results for learning representations for many applications, e.g. , images, speech [Oord *et al.*, 2018], and text [Mikolov *et al.*, 2013]. We are the first to explore its application for learning fair representations. Contrastive learning has been most actively explored in self-supervised learning, where the information to optimize is chosen by hand to be similar to some target task [Chen *et al.*, 2020; Oord *et al.*, 2018]. In our work, we demonstrated a natural connection between parity and mutual information. Other variational bounds on information [Poole *et al.*, 2019] and estimators like MINE [Belghazi *et al.*, 2018] and NWJ [Nguyen *et al.*, 2010] could also be leveraged for parity control using our results.

6 Conclusion

Most of the existing methods do not provide a way to control parity, and even if they do, often, it is only in a heuristic way. By proving a one-to-one relationship between information-theoretic quantities and statistical parity of arbitrary classifiers, we can finally see how varying a single hyper-parameter controlling information can explore the entire fairness versus accuracy spectrum. This information-theoretic characterization is algorithm-independent so that our control of parity can be guaranteed regardless of downstream applications.

References

[Alemi *et al.*, 2017] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 531–540, 2018.

[Bird *et al.*, 2020] Sarah Bird, Miroslav Dudík, Hanna Wallach, and Kathleen Walker. Fairlearn : A toolkit for assessing and improving fairness in AI. Technical report, 2020.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[Cisse and Koyejo, 2019] Moustapha Cisse and Sanmi Koyejo. Fairness and representation learning. Technical report, 2019.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Dutta *et al.*, 2019] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush R Varshney. An information-theoretic perspective on the relationship between fairness and accuracy. *arXiv preprint arXiv:1910.07870*, 2019.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 2012.

[Edwards and Storkey, 2016] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

[Higgins *et al.*, 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

[Jaiswal *et al.*, 2020] Ayush Jaiswal, Daniel Moyer, Greg Ver Steeg, Wael AbdAlmageed, and Premkumar Natarajan. Invariant representations through adversarial forgetting. In *AAAI*, pages 4272–4279, 2020.

[Louizos *et al.*, 2016] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The Variational Fair Autoencoder. In *International Conference on Learning Representations*, 2016.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, pages 3384–3393, 2018.

[McNamara *et al.*, 2017] Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

[Menon and Williamson, 2018] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pages 107–118, 2018.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, 2013.

[Moyer *et al.*, 2018] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, volume 31, pages 9084–9093, 2018.

[Nguyen *et al.*, 2010] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[Oord *et al.*, 2018] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[Poole *et al.*, 2019] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

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

[Song *et al.*, 2019] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2164–2173, 2019.

[Tange, Ole, 2020] Tange, Ole. GNU Parallel 20200522 ('Kraftwerk'), May 2020. GNU Parallel is a general parallelizer to run multiple serial command line programs in parallel without changing them.

[Xu *et al.*, 2020] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, pages 325–333, 2013.