A MAXIMAL CORRELATION APPROACH TO IMPOSING FAIRNESS IN MACHINE LEARNING

Joshua Lee[⋆†] *Yuheng Bu*[⋆†] *Prasanna Sattigeri*[‡] *Rameswar Panda*[‡]
Gregory Wornell[†] *Leonid Karlinsky*[‡] *Rogerio Feris*[‡]

[†] Department of EECS, MIT [‡] MIT-IBM Watson AI Lab, IBM Research

ABSTRACT

As machine learning algorithms grow in popularity and diversify to many industries, ethical and legal concerns regarding their fairness have become increasingly relevant. We explore the problem of algorithmic fairness, taking an information-theoretic view. The maximal correlation framework is introduced for expressing fairness constraints and shown to be capable of deriving regularizers that enforce independence and separation-based fairness criteria, which admit optimization algorithms that are more computationally efficient than existing algorithms. We show that these algorithms provide smooth performance-fairness tradeoff curves and perform competitively with state-of-the-art methods on the Communities and Crimes dataset.

***Index Terms*—** Fairness, HGR maximal correlation

1. INTRODUCTION

The use of machine learning in many industries has raised numerous ethical and legal concerns, including fairness and bias in predictions [1]. As systems are trusted to aid or make decisions regarding loan approval, criminal sentencing, and even health care, it is vital that unfair biases do not influence them. However, mitigating these biases is complicated by ever-changing perspectives on fairness, and a good system for enforcing fairness must be adaptable to new settings. In particular, there are often competing notions on fairness. Two popular notions are independence and separation, as discussed in [2]. Previous work, including [2], has proven that independence and separation are inherently incompatible for non-trivial cases and their applicability needs to be determined by the application and the stakeholders. This motivates us to construct a framework that is flexible enough to handle different fairness criteria.

This bias mitigation must also be balanced out with the system's usefulness, and often one must tune the tradeoff between the fairness (as measured based on the context) and performance according to current needs, which can be difficult if the tradeoff curve is not smooth. Generating the frontier of possible values can be computationally infeasible or impossible if the algorithm does not have a regularization parameter to adjust (see, [3, 4]), which makes fast generation of fair classifiers even more important.

Different contexts also require different points of intervention during the learning process to ensure fairness. *Pre-processing* ([3, 5]) approaches modify the data to eliminate bias whereas *post-processing* ([6, 7]) modify learned features/predictions from existing models to be more fair. We focus on the *in-processing* approach ([8, 4]), where fairness criteria are directly incorporated into the training objective to produce fairer learned features. Motivated by few-shot applications where only a pre-trained network and few samples labeled with the sensitive attribute are available, we also seek a method that is applicable in a post-processing manner when we have access to only a small number of samples labeled with the sensitive attribute that we wish to be fair about, which would arise in settings where collecting this information can be very difficult.

As existing approaches can struggle with efficiency, can fail to provide good control over the performance-fairness tradeoff, and/or can only deal with discrete variables. In this paper, we make the following contributions:

We present a framework justified by an information-theoretic view that can inherently handle the popular fairness criteria, namely independence and separation, which can be applied to continuous labels and sensitive attributes, and which uses the maximal correlation to construct measures of fairness associated with different criteria, then uses these measures to further develop fair learning algorithms in a fast, efficient, and effective manner.

We show empirically that these algorithms provide the desired smooth tradeoff curve between performance and fairness on the Communities and Crimes dataset.

Finally, we perform experiments to illustrate that our algorithms can be used to impose fairness on a model originally trained without any fairness constraint in the few-shot regime, which further demonstrates the versatility of our algorithms in a post-processing setup.

⋆Equal contribution. The work was done while J. Lee was at MIT; the author is now with Snap. This work was supported, in part, by the MIT-IBM Watson AI Lab under Agreement No. W1771646, and NSF under Grant No. CCF-1717610.

2. BACKGROUND

2.1. Fairness Objectives in Machine Learning

Consider the standard supervised learning scenario where we predict the value of a target variable $Y \in \mathcal{Y}$ using a set of decision or predictive variables $X \in \mathcal{X}$ with training samples $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. For example, X may be information about an individual's credit history, and Y is whether the individual will pay back a certain loan. In general, we wish to find features $f(x)$, which are predictive of Y, so that we can construct a good predictor $\hat{y} = T(f(x))$ of y under some loss criteria $L(\hat{y}, y)$.

Now suppose we have some sensitive attributes $D \in \mathcal{D}$ we wish to be "fair" about (e.g. race, gender), and training samples $\{(x_1, y_1, d_1), \ldots, (x_n, y_n, d_n)\}$. For example, in the criminal justice system, predictions about the chance of recidivism of a convicted criminal (Y) given factors such as the nature of the crime and the number of prior arrests (X) should not be determined by race (D). This is a known issue with the COMPAS recidivism score, which, despite not using race as an input to make decisions, still leads to systematic bias towards members of certain races in the output score [9].

The two most popular criteria for fairness are independence and separation. Independence states that for a feature to be fair, it must satisfy the independence property $\hat{Y} \perp D$ or $f(x) \perp D$. The intuition is simple: if the prediction/feature is independent of the sensitive attribute, then no information about the sensitive attribute is used to predict Y. This criterion has been studied under the lens of *demographic parity* and *disparate impact* in [2], and admits a class of fairness measures based on the degree of dependence between $f(X)$ and D. For example, independence is satisfied if and only if the mutual information $I(f(X); D)$ is zero.

Separation requires the conditional independence property $(\hat{Y} \perp D)|Y$ or $(f(X) \perp D)|Y$. This criterion allows for violation of demographic parity to the extent that it is justified by the target variable. In the general case, this criterion suggests a fairness measure based on the conditional dependence between $\hat{Y}$ and D conditioned on Y. For a more complete discussion of the advantages and disadvantages of these two criteria, please refer to [2].

2.2. Maximal Correlation

Since these fairness criteria are expressed as enforcing independencies with respect to joint distributions, we look for constraints that reduce dependency between variables. In particular, the right formulation of correlation between learned features and sensitive attributes can provide a framework for measuring and optimizing for fairness. One effective measure applicable to both continuous and discrete data is the Hirschfeld-Gebelein-Renyi (HGR) maximal correlation, a measure of nonlinear correlation which originated in [10] and is further developed in [11, 12]. The HGR maximal correlation between two random variables is equal to zero if and only if the two variables are independent, and increases in value the more correlated they are (i.e., the more biased/unfair).

Definition 1 *For two jointly distributed random variables $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$, given $1 \leq k \leq K - 1$ with $K = \min\{|\mathcal{X}|, |\mathcal{Y}|\}$, the HGR maximal correlation problem is*

$$(\mathbf{f}^*, \mathbf{g}^*) \triangleq \underset{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^k,\ \mathbf{g}\colon \mathcal{Y}\to\mathbb{R}^k}{\arg\max} \ \mathbb{E}\left[\mathbf{f}^{\mathrm{T}}(X)\,\mathbf{g}(Y)\right], \tag{1}$$

with constraints $\mathbb{E}[\mathbf{f}(X)] = \mathbb{E}[\mathbf{g}(Y)] = \mathbf{0}$, $\mathbb{E}\left[\mathbf{f}(X)\mathbf{f}^{\mathrm{T}}(X)\right] = \mathbb{E}\left[\mathbf{g}(Y)\mathbf{g}^{\mathrm{T}}(Y)\right] = \mathbf{I}$, and expectations taken over $P_{X,Y}$. We refer to $\mathbf{f}^* = (f_1^*, \ldots, f_k^*)^{\mathrm{T}}$ and $\mathbf{g}^* = (g_1^*, \ldots, g_k^*)^{\mathrm{T}}$ as maximal correlation functions, and the associated maximal correlations are

$$\sigma_i \triangleq \mathbb{E}\left[f_i^*(X)\, g_i^*(Y)\right], \text{ for } i = 1, \ldots, k, \tag{2}$$

and the HGR maximal correlation is $\mathrm{HGR}_k(X, Y) \triangleq \sum_{i=1}^k \sigma_i$.

Note that the original definition of HGR maximal correlation is the special case of our definition when $k = 1$ (see, [13]). This generalization of maximal correlation analysis enables us to produce more than one feature mapping by solving the maximal correlation problem, and these feature mappings can be used in other applications, including ensemble learning, multi-task learning, and transfer learning [14, 15].

The HGR is also linked to the mutual information via the following approximation, which holds when the joint distribution of X and Y are close to being independent [13]:

$$I(X; Y) \approx \frac{1}{2} \sum_{i=1}^{K-1} \sigma_i^2. \tag{3}$$

2.3. Related Work

Independence and separation have been studied in many works. Most existing approaches fail to provide an efficient solution in the continuous settings [6]. Other methods can also be limited in their ability to handle all dependencies between variables. Zafar et al. [16] uses a covariance-based constraint to enforce fairness, so it likely would not do well on other metrics. Furthermore, it is strictly a linear penalty rather than our non-linear formulation and penalizes the predictions of the system rather than the features learned. This limits the relationships between variables it can capture.

Mary et al. [4] propose the use of the HGR maximal correlation as a regularizer for either the independence or the separation constraint. In contrast to our approach dealing with the maximal correlation directly, they use a χ^2 divergence computed over a mesh grid to upper bound the HGR maximal correlation during the optimization of the classifier (either a linear regressor or a Deep Neural Net (DNN)). This method applies to cases where X is continuous and Y and D are either continuous or discrete variables, but scales poorly with the bandwidth and dimensionality of D.

There are other works which use either an HGR-based or mutual information-based formulation of fairness, but do not generalize to more than one setting. Grari et al. [17] and Baharlouei et al. [18] use correlation-based regularizers, but can only be used in the independence case. Furthermore, [18] only uses a single mode of the HGR maximal correlation (as opposed to multiple modes used here) for regularization, which limits the information it can encapsulate, and is also not designed for continuous sensitive attributes. Moyer et al. [19] also develops a method which can only be used for independence, and requires training an additional network in order to evaluate a bound for the mutual information which can be used to as a fairness penalty, thus increasing the complexity and required runtime. Finally, Cho et al. [20] approximates the mutual information with a variational formulation, but does not include a formulation for continuous labels.

3. MAXIMAL CORRELATION FOR FAIRNESS

Equipped with the HGR maximal correlation as a measure of dependence, we explore its use as a fairness penalty in the case where X, Y, and D are all continuous and real-valued.

In order to make the space of maximal correlation functions tractable, we first limit our scope of learning algorithms to those which train models (e.g. neural nets) via gradient descent using samples, which encompasses most commonly-used methods. We thus restrict the space of maximal correlation functions to be the family of functions that can be learned by neural nets, allowing us to compute the gradient while still providing a rich set of functions to search over.

3.1. Independence

To ensure sufficient independence, we want to minimize the loss function $L(\hat{Y}, Y)$ and the maximal correlation between $\mathbf{f}(X)$ and D. Our optimization (for a given λ) then becomes:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} L(T(\mathbf{f}(X)), Y) + \lambda \mathrm{HGR}_k(\mathbf{f}(X), D), \quad (4)$$

where $\mathrm{HGR}_k(\mathbf{f}(X), D) = \max_{\mathbf{g},\,\mathbf{h}} \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\,\mathbf{h}(D)\right]$, $\mathbb{E}\left[\mathbf{g}(\mathbf{f}(X))\right] = \mathbb{E}\left[\mathbf{h}(D)\right] = \mathbf{0}$, and $\mathbb{E}\left[\mathbf{g}(\mathbf{f}(X))\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\right] = \mathbb{E}\left[\mathbf{h}(D)\mathbf{h}^{\mathrm{T}}(D)\right] = \mathbf{I}$. m is the dimension of the features $\mathbf{f}(X)$, k is the number of maximal correlation functions, and $\mathbf{g}\colon \mathbb{R}^m \to \mathbb{R}^k$, $\mathbf{h}\colon \mathcal{D} \to \mathbb{R}^k$ are the maximal correlation functions relating $\mathbf{f}(X)$ with D. Given the difficulty of enforcing the orthogonalization constraint, we use a variational characterization of the HGR maximal correlation called Soft-HGR proposed in [15] which relaxes the orthogonal constraint:

$$\mathrm{HGR}_{\mathrm{soft}}(X, Y) \triangleq$$

$$\max_{\substack{\mathbb{E}[\mathbf{g}(X)]=\mathbf{0} \\ \mathbb{E}[\mathbf{h}(Y)]=\mathbf{0}}} \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(X)\,\mathbf{h}(Y)\right] - \frac{1}{2}\,\mathrm{tr}(\mathrm{cov}[\mathbf{g}(X)]\,\mathrm{cov}[\mathbf{h}(Y)]),$$

where $\mathrm{cov}[X]$ is the covariance matrix of X. [15] shows that this Soft-HGR formulation can be viewed as a low-rank approximation of the original HGR maximal correlation problem in the discrete case. Then,our learning objective becomes:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} \max_{\substack{\mathbf{g}\colon \mathbb{R}^m\to\mathbb{R}^k,\ \mathbf{h}\colon \mathcal{D}\to\mathbb{R}^k \\ \mathbb{E}[\mathbf{g}(\mathbf{f}(X))]=\mathbb{E}[\mathbf{h}(D)]=\mathbf{0}}} C, \quad (5)$$

$$\text{with} \quad C = L(T(\mathbf{f}(X)), Y) + \lambda \mathbb{E}\left[\mathbf{g}^{\mathrm{T}}(\mathbf{f}(X))\,\mathbf{h}(D)\right] - \frac{\lambda}{2}\,\mathrm{tr}\left(\mathrm{cov}[\mathbf{g}(\mathbf{f}(X))]\,\mathrm{cov}[\mathbf{h}(D)]\right).$$

We solve this optimization by alternating between optimizing $\mathbf{f}, T$ and optimizing $\mathbf{g}, \mathbf{h}$. In practice, we implement this by alternating between one step of gradient descent for $\mathbf{f}$ and T and 5 steps of gradient descent on $\mathbf{g}$ and $\mathbf{h}$ to allow the maximal correlation functions to adapt to the changing of $\mathbf{f}$.

3.2. Separation

For the separation criterion, we want to ensure sufficient conditional independence $(f(X) \perp D)|Y$. Since maximal correlation is related to mutual information, we consider the following formulation:

$$I(\mathbf{f}(X); D, Y) = I(\mathbf{f}(X); Y) + I(\mathbf{f}(X); D|Y) \quad (6)$$

where the equality follows from the chain rule of mutual information. Thus, we can control the conditional mutual information $I(\mathbf{f}(X); D|Y)$, by using $I(\mathbf{f}(X); D, Y) - I(\mathbf{f}(X); Y)$ as a regularizer in the training process.

By replace the mutual information terms with the maximal correlation terms, our optimization problem becomes:

$$\min_{\substack{\mathbf{f}\colon \mathcal{X}\to\mathbb{R}^m \\ T\colon \mathbb{R}^m\to\mathcal{Y}}} L(T(\mathbf{f}(X)), Y) + \lambda(\mathrm{HGR}_{\mathrm{soft}}(f(X), (D \times Y)) - \mathrm{HGR}_{\mathrm{soft}}(f(X), Y)). \quad (7)$$

Note that for the first soft-HGR term, we use $\mathbf{g}, \mathbf{h}$ to denote the maximal correlation functions, and $\mathbf{g}', \mathbf{h}'$ to denote the functions for the second term. Once again, we solve this optimization by alternating between optimizing $\mathbf{f}$, T and optimizing $\mathbf{g}, \mathbf{h}, \mathbf{g}', \mathbf{h}'$.

3.3. Few-shot Learning

Our learning objective can also be applied *a posteriori* in a few-shot setting with a classifier that has already been trained in a fairness-unaware manner on a large number of samples without the sensitive attribute label. In this case, we can formulate our objective as before, and use the few samples containing the sensitive attribute to further train the network and force it to learn fairer features that are still predictive of the desired labels.

4. EXPERIMENTAL RESULTS

In order to illustrate the effectiveness of our algorithms, we experiment on the Communities and Crimes (C&C) dataset[1].

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

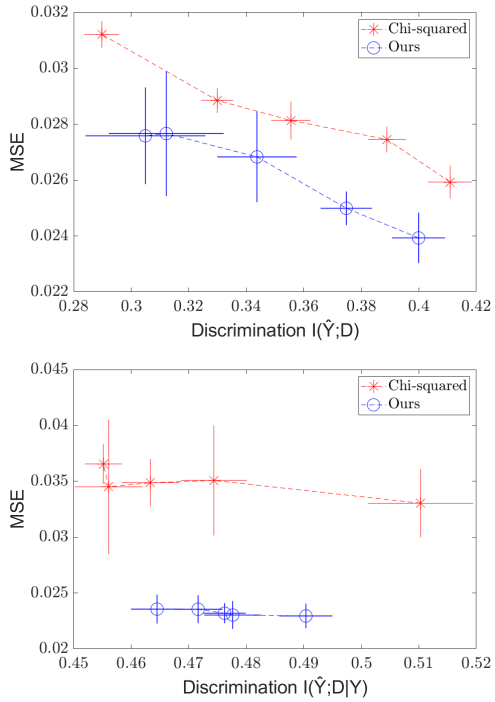

Fig. 1. Independence (top) and Separation (bottom) result on the C&C dataset.

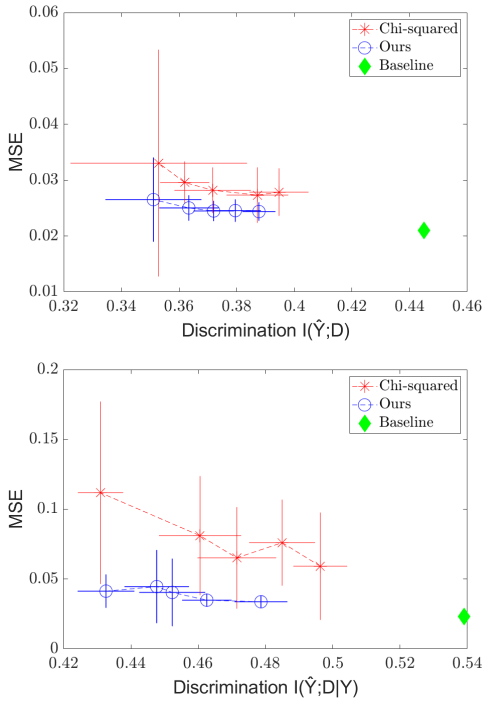

Fig. 2. Independence (top) and Separation (bottom) result on the C&C dataset in the *few-shot* settings.

The goal is to predict the crime rate Y of a community given a set of 121 statistics X (distributions of income, age, urban/rural, etc.). The 122-th statistic (percentage of black people in the community) is used as the sensitive variable D. All variables in this dataset are real-valued. The dataset was split into 1794 training and 200 test samples. Following [4], we use a Neural Net with a 50-node hidden layer (which we denote as $f(x)$) and train a predictor $\hat{y} = T(f(x))$ with the mean squared error (MSE) loss and the soft-HGR penalty, varying λ. For soft-HGR, we use two 2-layer NNs with scalar outputs as the two maximal correlation functions $\mathbf{g}$ and $\mathbf{h}$, and trained them according to (5) (independence) or (7) (separation). We then computed the test MSE and test "discrimination" in each case.

For independence, our metric was $I(\hat{Y}; D)$, approximated using a standard kNN-based mutual information estimator [21]. For separation, we computed $I(\hat{Y}; D|Y)$ with the same estimator. We report the results of our experiment as well as that of the χ^2 method of [4] with the same architecture. The results of the experiments are presented in Figure 1.

As expected, we see a tradeoff between the MSE and discrimination, creating a frontier of possible values. We also see that the Soft-HGR penalty provides gains compared to the χ^2 method for both independence and separation. Moreover, our method runs significantly faster than the χ^2 method (on the order of seconds per iteration for our method versus just under a minute per iteration for the comparison method), as the χ^2 method requires computation over a mesh grid of a Gaussian KDE, which scales with the product of the number of "bins" (mesh points) and the number of training samples, while our method only scales with the number of samples ($O(n)$). KDE methods also scale poorly with dimensionality (see, [22]) in an exponential manner, and thus if d is high-dimensional, the χ^2 method would run much slower than our method, which can take in an arbitrarily-sized input and scale linearly with the dimensionality of the input.

We also run experiments to illustrate how our method's simplicity allows it to adapt to the few-shot, few-epoch regime faster than that of the χ^2 method. We take 10 "few-shot" samples from the training set, then train a network to predict Y from X *without* any fairness regularizer using the full training set. Then, we run 5 more iterations of gradient descent on the trained model using the fairness-regularized objective and the 10 few-shot samples, and compare the results between the Soft-HGR and χ^2 regularizer. We choose to compare to the χ^2 regularizer as it is one of the few methods designed to handle continuous D. The results are shown in Figure 2. Once again, we see the tradeoff curve, and see our method outperform the χ^2 method, and that it appears to be competitive with the standard case in just a few iterations, while the χ^2 method is still far from achieving the original MSE. We also vastly outperform the baseline (before fairness regularization) model in reducing discrimination, at the cost of only a small increase in error. Thus, in situations where only a few samples labeled with the sensitive attribute can be collected, fairness can still be enforced.

5. REFERENCES

[1] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 59–68.

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019, `http://www.fairmlbook.org`.

[3] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[4] Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *International Conference on Machine Learning*, 2019, pp. 4382–4391.

[5] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney, "Fairness gan," *arXiv preprint arXiv:1805.09910*, 2018.

[6] Moritz Hardt, Eric Price, and Nathan Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems 29*, Barcelona, Spain, Dec. 2016, pp. 3315–3323.

[7] Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon, "Optimized score transformation for fair classification," *arXiv preprint arXiv:1906.00066*, 2019.

[8] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[9] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks," *ProPublica*, 2016.

[10] Hermann O. Hirschfeld, "A connection between correlation and contingency," *Proc. Cambridge Phil. Soc.*, vol. 31, pp. 520–524, 1935.

[11] Hans Gebelein, "Das statistische problem der korrelation als variations-und eigenwertproblem und sein zusammenhang mit der ausgleichsrechnung," *Z. Angewandte Math., Mech.*, vol. 21, no. 6, pp. 364–379, 1941.

[12] Alfréd Rényi, "On measures of dependence," *Acta Mathematica Academiae Scientiarum Hungarica*, vol. 10, no. 3–4, pp. 441–451, Sept. 1959.

[13] Shao-Lun Huang, Anuran Makur, Gregory W Wornell, and Lizhong Zheng, "On universal features for high-dimensional learning and inference," *arXiv preprint arXiv:1911.09105*, 2019.

[14] Joshua Lee, Prasanna Sattigeri, and Gregory Wornell, "Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks," in *Advances in Neural Information Processing Systems*, 2019, pp. 4372–4382.

[15] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 5281–5288.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 962–970.

[17] Vincent Grari, Boris Ruf, Sylvain Lamprier, and Marcin Detyniecki, "Fairness-aware neural Rényi minimization for continuous features," *arXiv preprint arXiv:1911.04929*, 2019.

[18] Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn, "Rènyi fair inference," *arXiv preprint arXiv:1906.12005*, 2019.

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

[20] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh, "A fair classifier using mutual information," in *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2020, pp. 2521–2526.

[21] Weihao Gao, Sewoong Oh, and Pramod Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5629–5661, 2018.

[22] Zhipeng Wang and David W Scott, "Nonparametric density estimation for high-dimensional data—algorithms and applications," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 4, pp. e1461, 2019.