Fair machine learning

Lecture 1

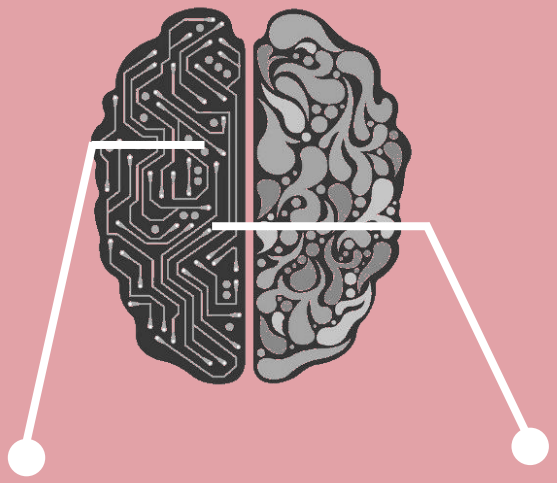
Changho Suh
EE, KAIST

Aug. 3, 2021

Overview

Reading: **Tutorial Note (TN) 1**

AI



Google assistant

Self driving

recruiting

judgement

loan decision

Trustworthy AI

"AI has significant potential to help solve challenging problems, including by advancing medicine, understanding language, and fueling scientific discovery. ***To realize that potential****, it's critical that* ***AI*** *is used and developed* ***responsibly."***

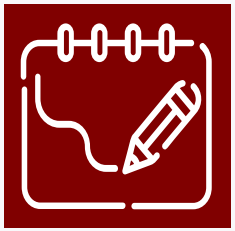
"Moving forward, "build for performance" will not suffice as an AI design paradigm. We must learn how to build, evaluate and monitor for ***trust****."*

Five aspects of trustworthy AI

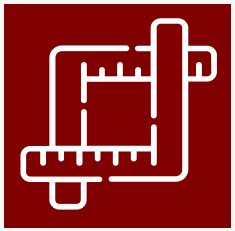
focus of this tutorial

fairness

robustness



explainability



value alignment

transparency

Two contexts of this tutorial's focus

Will explore **fairness** issues in the following two contexts:

1. **Fair classifiers**

   Role: Make unbiased decisions across different groups and/or individuals.

2. **Fair generative models**

   Role: Generate fake samples that resemble real data while ensuring fairness.

Fair classifiers

Fairness in the context of classifers

There are many fairness concepts.

One important concept is group fairness:

Pursues predictions to exhibit similar statistics regardless of sensitive attributes of groups

e.g., race, gender, age, religion, etc.

Applications of fair classifiers

job hiring

parole decision

Applicants want no discrimination depending on race or sex.

A fair predictor for recidivism score (e.g., reoffending probability) plays a crucial role.

A fairness measure

Feldman et al. SIGKDD15

Zafar et al. AISTATS17

Y: class $\in \{0, 1\}$ $\tilde{Y}$: prediction (hard decision)

no reoffend, reoffend

black, white

Z: sensitive attribute e.g., $\in \mathcal{Z} = \{0, 1\}$

Demographic Parity (DP) condition:

$$\tilde{Y} \perp Z: \quad \mathbb{P}(\tilde{Y} = 1 | Z = z) = \mathbb{P}(\tilde{Y} = 1), \forall z \in \mathcal{Z}$$

A quantifed measure: **D**ifference btw two interested probabilities in **DP** condition

$$\text{DDP} := \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1 | Z = z) - \mathbb{P}(\tilde{Y} = 1)|$$

Limitation of DP condition

Demographic Parity (DP) condition:

$$\tilde{Y} \perp Z: \quad \mathbb{P}(\tilde{Y}=1|Z=z) = \mathbb{P}(\tilde{Y}=1), \forall z \in \mathcal{Z}$$

Suppose that the ground truth distribution respects:

$$\mathbb{P}(Y=1|Z=1) \gg \mathbb{P}(Y=1|Z=0)$$

Enforcing the DP condition may aggravate prediction accuracy significantly.

Another fairness notion

Hardt-Price-Srebo NeurIPS16

Equalized Odds (EO) condition: $\tilde{Y} \perp Z \,|Y$

$$\mathbb{P}(\tilde{Y} = 1|Y = y, Z = z) = \mathbb{P}(\tilde{Y} = 1|Y = y) \;\forall z \in \mathcal{Z}, \forall y \in \mathcal{Y}$$

Enforcing the EO condition may not necessarily reduce prediction accuracy.

A quantified measure:

$$\text{DEO} := \sum_{y \in \mathcal{Y}} \sum_{z \in \mathcal{Z}} |\mathbb{P}(\tilde{Y} = 1|Y = y, Z = z) - \mathbb{P}(\tilde{Y} = 1|Y = y)|$$

Many recent works on fair classifiers

Here is only a *partial* list:

[Feldman et al. SIGKDD15]

[Hardt-Price-Srebo NeurIPS16]

[Pleiss et al. NeurIPS17]

[Zhang et al. AIES18]

[Donini et al. NeurIPS18]

[Agarwal et al. ICML18]

[Roh-Lee-Whang-**Suh** ICLR 21]

[Zafar et al. AISTATS17]

[Cho-Hwang-**Suh** ISIT20]

[Roh-Lee-Whang-**Suh** ICML20]

[Cho-Hwang-**Suh** NeurIPS20]

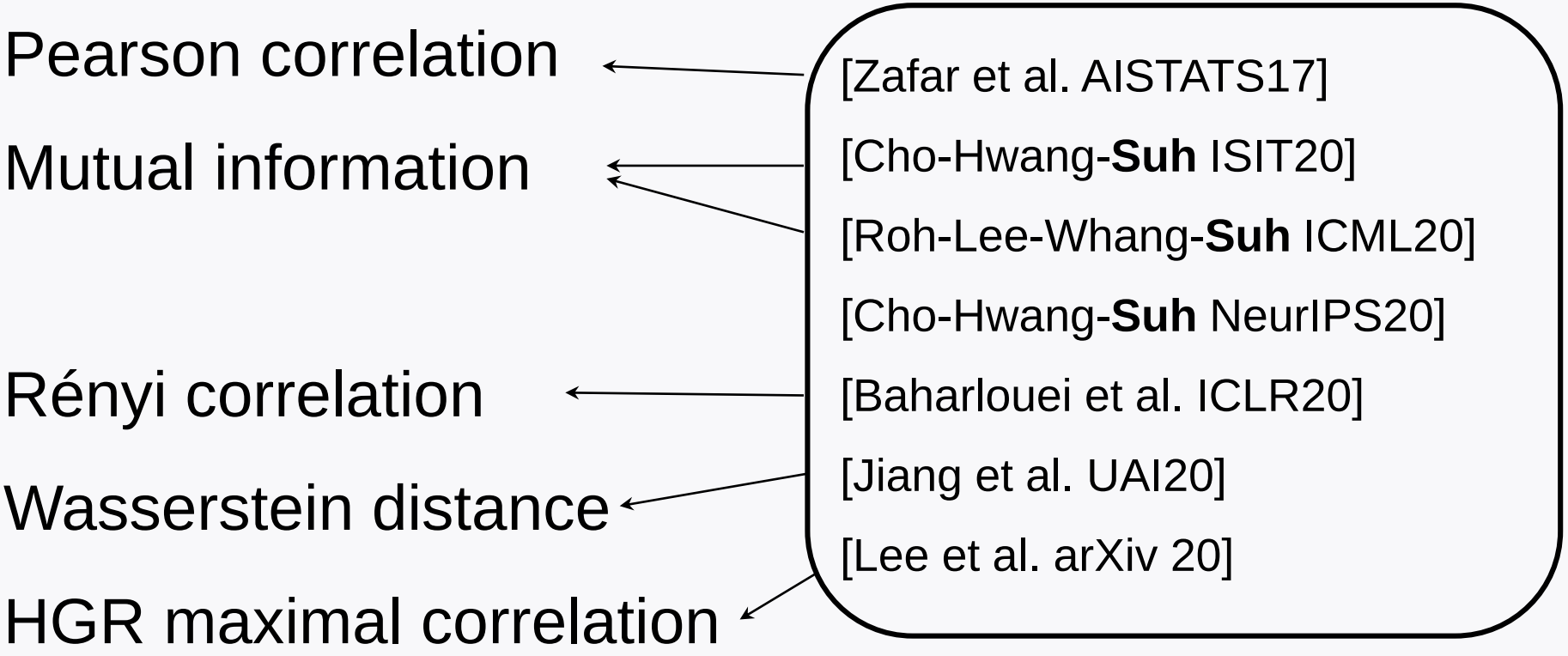
[Baharlouei et al. ICLR20]

[Jiang et al. UAI20]

[Lee et al. arXiv 20]

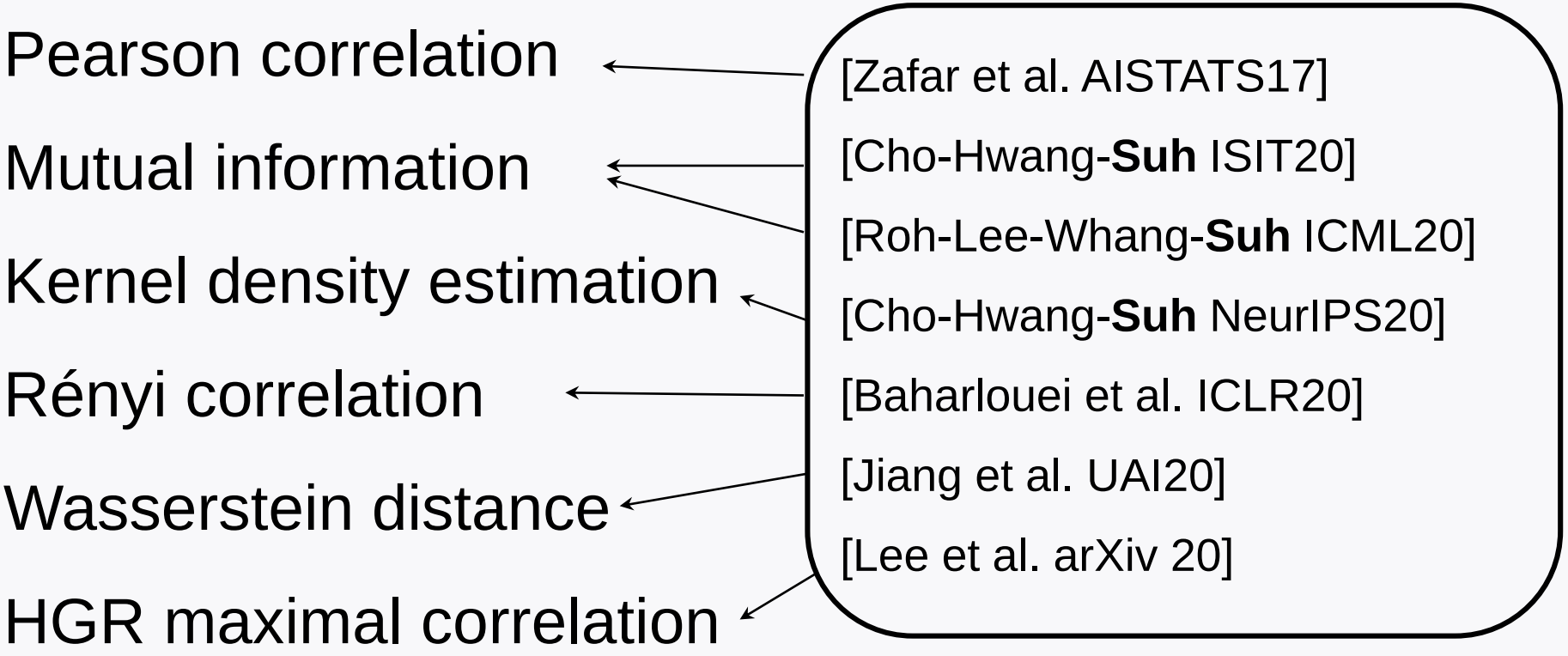
Approaches relevant to information theory

Employed measure:



Approaches relevant to information theory

Employed measure (or technique):



Focus of this tutorial

The most prominent information-theoretic measure

↓

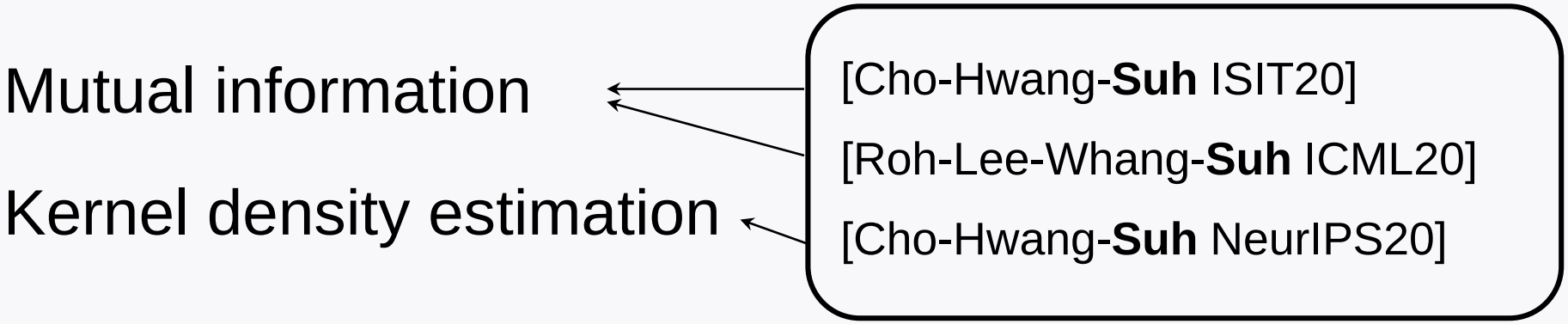
Mutual information

Kernel density estimation (KDE)

Turns out:

A KDE-based fair classifier is the state of the art.

Today's lectures



Lecture 2:

Will establish a connection btw fairness measure & mutual information.

Will study an MI-based fair classifier.

Lecture 3:

Explore a KDE-based fair classifier which offers a better accuracy-fairness tradeoff.

Fair generative models

Fairness in the context of generative models

There are many fairness concepts.

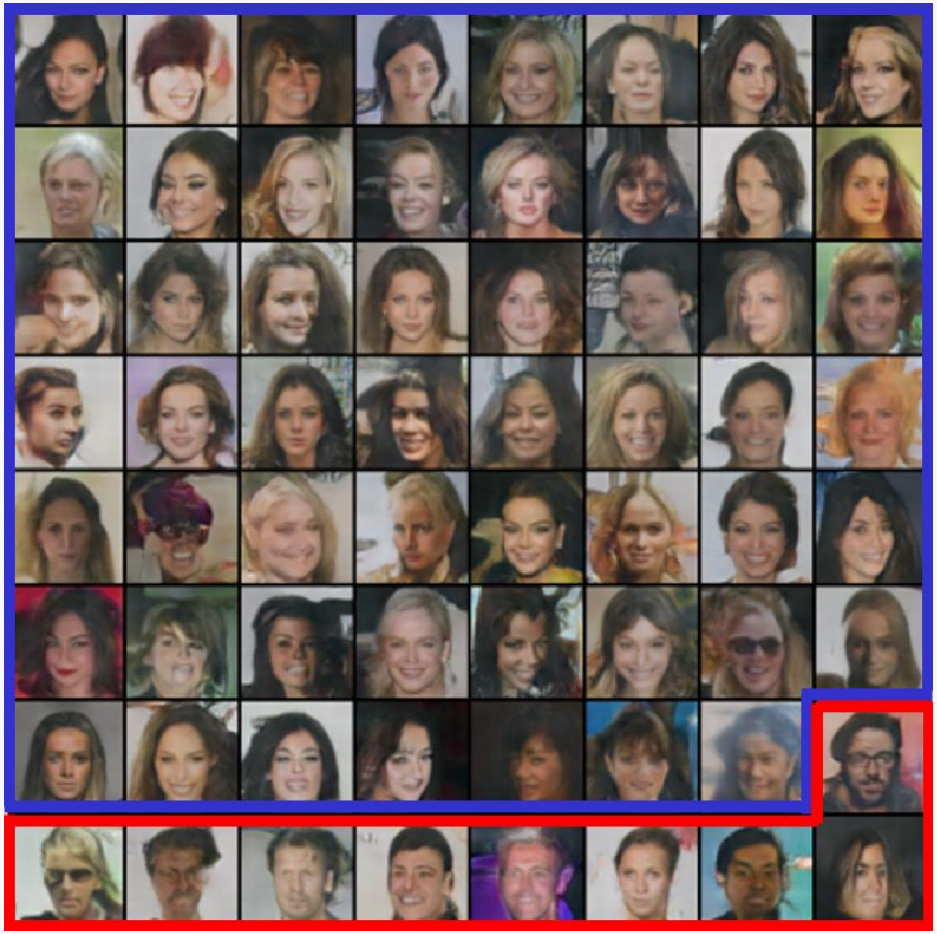
One important concept is **fair representation:**

> Pursues class-balanced generated samples even when trained with size-biased real data among different demographics.

Example: A partial “CelebA dataset” Liu et al. ICCV15

Female:male ~= 85:15

A naive generative model trained with this dataset yields *biased generated samples*:



Goal: Ensure **fair representation** even under such imbalanced real data.

A measure for fair representation Choi et al. ICML20

Fairness Discrepancy (FD):

$$\left\| \begin{bmatrix} \mathbb{P}_G(z_1) \\ \vdots \\ \mathbb{P}_G(z_{|\mathcal{Z}|}) \end{bmatrix} - \begin{bmatrix} \mathbb{P}_{\text{desired}}(z_1) \\ \vdots \\ \mathbb{P}_{\text{desired}}(z_{|\mathcal{Z}|}) \end{bmatrix} \right\|_2$$

desired dist. (e.g., uniform)

dist. of generated samples w.r.t. sensitive attribute Z

Recent works on fair generative models

[Xu et al. BigData18]

[Xu et al. BigData19]

[Sattigeri et al. IBM-Journal19]

[Jalal et al. ICML21]

[Yu et al. ECCV20]

[Choi et al. ICML20]

[Tan et al. arXiv20]

[Um-**Suh** ‘21]

Recent works on fair representation

[Yu et al. ECCV20]

[Choi et al. ICML20]

[Tan et al. arXiv20]

[Um-**Suh** ‘21]

Focus of this tutorial

[Um-**Suh** ’21]

Employs a well-known statistical measure that often arises in information theory:

Total Variation Distance (TVD)

Also: **State of the art**

Tomorrow's lectures

[Um-**Suh** '21]

TVD-based
State of the art

Note: Built upon a GAN variant based on hinge loss.

Lecture 4: Study the hinge-loss-based GAN.
Make a connection with TVD.

Lecture 5: Explore a TVD-based fair generative model.

Lecture 6: Discuss a couple of other relevant issues.

Look ahead

Embark on fair classifiers and explore a connection between fairness measure & mutual information.

Reference: fair classifiers

[1] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.

[2] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. *Artificial Intelligence and Statistics Conference (AISTATS)*, 2017.

[3] M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *In Advances in Neural Information Processing Systems 29 (NeurIPS)*, 2016.

[4] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. *In Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[5] B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2018.

Reference: fair classifiers

[6] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. *In Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

[7] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. *In Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[8] Y. Roh, K. Lee, S. E. Whang and C. Suh. FairBatch: Batch selection for model fairness. *International Conference on Learning Representations* (*ICLR),* 2020.

[9] J. Cho, G. Hwang and C. Suh. A fair classifier using mutual information. *IEEE International Syposium on Inofrmation Theory* (*ISIT),* 2020.

[10] Y. Roh, K. Lee, S. E. Whang and C. Suh. FR-Train: A mutual information-based approach to fair and robust training. *In Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

Reference: fair classifiers

[11] J. Cho, G. Hwang and C. Suh. A fair classifier using kernel density estimation. *In Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[12] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. Renyi fair inference. *International Conference on Learning Representations* (*ICLR),* 2020.

[13] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein Fair Classification. *In Proceedings of the 35th Uncertainty in Artificial Intelligence Conference* (*UAI),* 2020.

[14] J. Lee, Y. Bu, P. Sattigeri, R. Panda, G. Wornell, L. Karlinsky, and R. Feris. A maximal correlation approach to imposing fairness in machine learning. *arXiv:2012.15259*, 2020.

Reference: fair generative models

[1] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. *IEEE International Conference on Big Data (Big Data)*, 2018.

[2] D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. *IEEE International Conference on Big Data (Big Data)*, 2019.

[3] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 2019.

[4] A. Jalal, S. Karmalkar, J. Hoffmann, A. G. Dimakis, and E. Price. Fairness for Image Generation with Uncertain Sensitive Attributes. *International Conference on Machine Learning (ICML)*, 2021.

[5] N. Yu, K. Li, P. Zhou, J. Malik, L. Davis, and M. Fritz. Inclusive gan: Improving data and minority coverage in generative models. *European Conference on Computer Vision (ECCV)*, 2020.

Reference: fair generative models

[6] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon. Fair generative modeling via weak supervision. *In Proceedings of the 37th International Conference on Machine Learning (ICML),* 2020.

[7] S. Tan, Y. Shen, and B. Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

[8] S. Um and C. Suh. A fair generative model using total variation distance. *submitted to Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

[9] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. *International Conference on Computer Vision (ICCV)*, December 2015.

[10] J. H. Lim and J. C. Ye. Geometric gan. *arXiv:1705.02894*, 2017.

[11] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. *International Conference on Learning Representations*, 2019.