Fair Latent Deep Generative Models (FLDGMs) for Syntax-agnostic and Fair Synthetic Data Generation - Supplementary Materials

1 Background

1.1 Distance correlation

State-of-the-art studies [2, 12] point that a dependence measure should satisfy seven properties such as symmetry, boundedness, monotonicity, etc. Both mutual information and distance correlation satisfy five properties out of seven mentioned in [2]. One problem with mutual information-based dependence measure is that it should need an adversary to approximate the upper bound, which is unstable. Distance correlation accounts for this and can better balance the fairness-quality tradeoff. We compared various dependency measures such as Maximum Mean Discrepancy (MMD), Adversarial Debiasing (Ad), and Mutual Information (MI) on accuracy and fairness (in terms of DP) in Figure 1. It is obvious from the figure that distance correlation is superior in balancing the accuracy-fairness tradeoff.

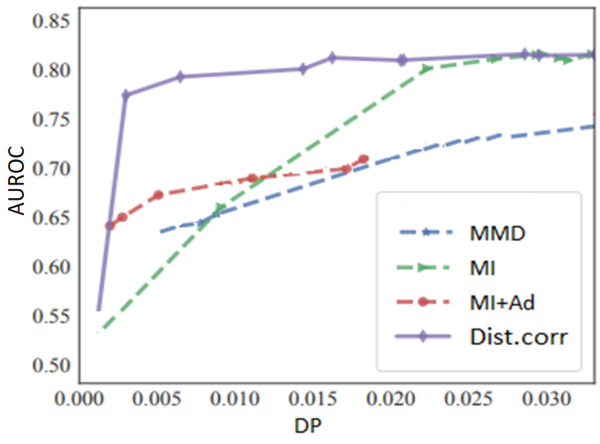

Figure 1: Comparison of different dependency measures on accuracy and fairness.

The complexity of computing distance correlation is dependent on the discrete values of S, latent space dimension, and batch size. In [8], the authors used matrix algebra and parallel computation to compute this fairness penalty.

2 Ablation study on the value of fairness penalty coefficient (α)

We analyze the performance in terms of accuracy and fairness with different α and set $\alpha = 7$ as it balances the quality-fairness tradeoff as shown in Figure 2.

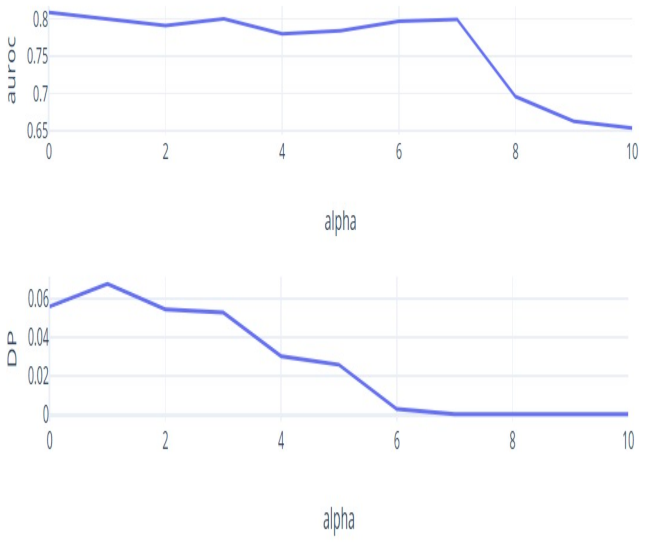

Figure 2: Ablation study on the value of α.

3 Theorems and proofs

Lemma 1. Given an optimal discriminator D in GAN and a fixed diffusion process in DM, the global minimum of $\mathbf{G}$ is achieved if and only if $p_{\mathbf{G}} = p_{\mathbf{R}}$, where $p_{\mathbf{G}}, p_{\mathbf{R}}$ respectively denote the generated latent distribution and ground truth fair latent distribution.

Proof. For GAN, we refer to Theorem 1 from [4]. The GAN min-max optimization is performed by updating the generator G and the discriminator D with corresponding losses. Note that we have not altered the GAN training in FLDGM. So, all the theoretical results for GANs can be transferred to FLDGM. Therefore, the global minimum of G in GAN-based FLDGM can be achieved if and only if $p_G = p_{\mathbf{R}}$. We could not find any theoretical convergence results for DMs, but if exists, can be transferred to FLDGM as we have made zero changes to the DM training. Therefore the global minimum of a generator neural network $\mathbf{G}$ in FLDGM is achieved if and only if $p_{\mathbf{G}} = p_{\mathbf{R}}$. We denote $\mathbf{G}$, as a general representation of a generator network in FLDGM.

Theorem 1. *Assume that (i) the data generation is Markov compatible with a pre-trained autoencoder, which is optimized for a combination of fairness loss and reconstruction loss, (ii) the neural networks involved in DGM have enough capacity, and (iii) the training of all the components of DGM is iterative until optimality is achieved, then for a well-optimized FLDGM, the generated fair latent distribution $p_{Z'}$ by the generator network $\mathbf{G}$ in $\mathcal{G}$ always converges to the ground-truth fair latent distribution p_Z.*

Proof for Theorem 1. Given the adequate capacity of $\mathbf{G}$, by the convex training of DGM and the existence of global optimum as stated in Lemma 1, the Z' always converges to Z.

Theorem 2. *For a well-optimized generative model $\mathcal{G}$ in FLDGM, the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ - fair, given the corresponding pre-trained autoencoder.*

Proof for Theorem 2. Let $\mathbf{G}$ be a generator neural network in $\mathcal{G}$. For a fixed $\mathbf{G}$, the $\mathcal{G}$ will converge to a true fair latent distribution Z and once optimized $\mathbf{G}$ can generate synthetic fair latent vectors Z' similar to Z (using Theorem 1). Thus the fairness, $\mathcal{U}(S, Y)$ contained in Z will be replicated in Z' for a well-optimized $\mathcal{G}$. Here, $\mathcal{U}(S, Y)$ is a definition of algorithmic fairness enforced in the fair abstract compression stage(autoencoder). Therefore the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ *- fair*.

Theorem 3. *Any optimal downstream classifiers* $\mathbf{M}$ *(without any explicit biases) trained on* $\mathbf{D}'$ *will have fair predictions on* $\mathbf{D}$ *under* $\mathcal{U}(S, Y)$.

Proof for Theorem 3. According to theorem 2, the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ *- fair*. Given the autoencoder and $\mathbf{G}$, we have access to the decoder $\mathcal{D}$. The transition from fair latent Z' to fair data $\mathbf{D}'$ can be done in a single pass through $\mathcal{D}$ without retraining. Suppose, we have an optimal downstream predictor $\mathbf{M}$, which can be any universal function approximator (e.g., MLP), that is trained on a sufficiently large quantity of the synthetic data $\mathbf{D}'$ generated by $\mathbf{G}$ and passed through $\mathcal{D}$, then the optimal prediction is $\mathcal{U}(S, Y)$ *- fair* as the decoder is pre-trained in stage 1 and we have not altered the training of the decoder while under reconstruction. Therefore, the definition of algorithmic fairness $\mathcal{U}(S, Y)$ will be satisfied by $\mathbf{D}'$.

Note that, we can achieve maximum fair performance when the model does not have any explicit biases. Therefore, for the sake of simplicity, we assume that the model does not add any biases during training.

4 Fairness analysis on Adult Income

4.1 Adult Income

For Adult income data, we used two fairness measures namely FTU and DP respectively for measuring direct and indirect discrimination. We selected 'gender' as a sensitive attribute as per studies [3] as there is a bias between 'gender' and 'income'. Around 68 percent of the gender population are men. One issue with this is that the model parameters may tend to be skewed toward the majority. For example, the trends will vary between the female and male populations. Trends mean the associations between attributes and the target variables. This may cause unfairness in the female population as the model maximizes accuracy across the entire population. In this case, we can define a privileged group as the male population and an unprivileged group as the female population.

Fairness Through Unawareness (FTU). It is used to analyze the direct influence of sensitive features, 'gender' in our case on income prediction. It can be calculated by the difference in predictions of a classifier for setting $'gender' = 1$ and $'gender' = 0$ (1 for male and 0 for female) such that the difference should be zero if the 'gender' has no direct influence on 'income'. We refer to Table 1 as the FTU for real data is 0.116 which is not fair. This means that the feature gender has a direct influence on income prediction. The metric FTU does account for direct discrimination as it only measures the direct influence.

Demographic Parity (DP). DP is based on the Predicted as Positive (PPP) rates. This means that DP measures the percentage of individuals who have either been correctly (TP) or incorrectly (FP) predicted as positive. In summary, this is the percentage of individuals who have benefited from the model prediction. Therefore, it measures the indirect influence of 'gender' on 'income' prediction. This is important in measuring all the features which contributed to the positive prediction, which means that if there is any other feature that has a strong correlation with 'gender', termed as a proxy attribute, it can also push some individuals to positive prediction. Therefore, it is a strong measure of indirect discrimination through proxy attributes. For a fair prediction, the DP should be close to zero, which means that both the unprivileged and privileged groups should have the same positive predictions. The DP of real data with S='gender' is 0.180 which is illegal.

5 Architecture and implementation details

5.1 Generative model architecture

In our Fair Latent Deep Generative Models (FLDGMs) framework, we used GAN-based and Diffusion-based architectures. For the GAN-based generative model, we used Least Square GAN [10] and Wasserstein GAN with gradient penalty [5] to generate fair latent space. For the generator and discriminator of both GAN architectures, we used a series of blocks consisting of batch normalization followed by linear layer and leaky relu activation functions. In the last layer, we used a linear layer.

For the diffusion architecture, we used the Gaussian diffusion model [6] which adds noise to the real data over some time steps T and makes the data a Gaussian noise in the forward process, and then in the reverse process, we train a neural network to approximate equation **??**. For the neural network architecture, we used two blocks consisting of batch normalization, linear layer, relu activation functions, and lastly a dropout layer. Also, to get the data in any timesteps $t \in T$, we used positional encoding before passing the data to the blocks. Figure 3 shows both the architecture for GAN and the diffusion model.

5.2 Autoencoder architecture

The autoencoder architecture we used consists of linear layer-ReLU-linear layer for both the encoder and decoder for Adult income data. For CMNIST, we used five pairs of Conv-ReLU with a flatten at the output for the encoder. For the decoder, we used five pairs of ConvTranspose2d-ReLU with a sigmoid at the output. We used a batch size of 2048 with a learning rate of 1e-3 and training epochs of 1000 for both data.

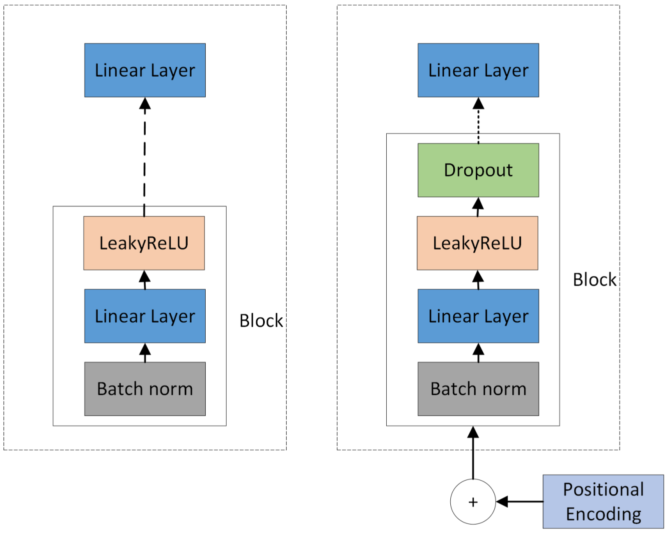

Figure 3: Architectures for GAN (left) and diffusion (right)

5.3 *Implementation details and Hyperparameters*

In order to generate fair latent space, first we need the ground truth. For that, we train both the Adult Income dataset and MNIST dataset with the autoencoder[8]. We followed the same hyperparameters used in the original experiment.

Once the ground truth fair latent space has been obtained, we trained GAN and diffusion models to generate a synthetic version of it. Hyperparameters for our models can be found in Table 1. In the FLD-WGAN-GP and FLD-LSGAN architectures, we used four blocks of network mentioned in Section 5 for the generator and three blocks for the discriminator. The layer dimension used in each block for FLD-WGAN-GP, FLD-LSGAN, and FLD-DM can be found in Table 2.

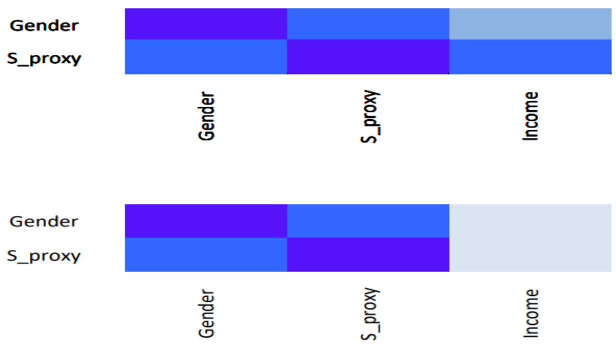

Figure 4: FLDGM accounts for proxy attributes (dark blue indicates high correlation)

6 Related work summary

An overview of related works in terms of different key areas of interest is given in Table 3.

7 Additional Results

7.1 *Comparison of conditional generation and FLDGM*

We conducted an additional experiment to see the difference between the fairness objectives based on conditional generation and distance correlation minimization. Conditional generation is done by generating balanced samples conditioned on the attribute ’gender’. In order to assess the effect of a proxy attribute, we created an extra feature S_{proxy} that is strongly correlated with ’gender’ in the Adult Income dataset. For the male sub-group, we set $S_{proxy} = 1$ for 95 percent of all cases, and for the remaining $S_{proxy} = 0$. For the female group, the above values are swapped. A correlation plot in Figure 4 shows that there is a strong correlation between S_{proxy} and income in the conditional generation. In FLDGM, this correlation has been minimized to a reasonable extent. Therefore the distance correlation-based fairness objective in FLDGM is superior in debiasing including proxy attributes, which proves the study in [14]. Also, conditional generation is not advisable in situations where there are more sensitive or proxy attributes, and keeping the balance between all these attributes is not easy during generation. Another line of research was pursued in [11], where the spurious correlation of proxy attributes is tackled by corrective sampling, which involves retraining the generative model followed by a discriminator rejection sampling [1].

7.2 *Prediction on CMIST data*

The CMNIST data [7] The classification performance has been improved in FLDGM which substantiates the importance of debiasing CMNIST data. An example of a prediction is given in Figure 5.

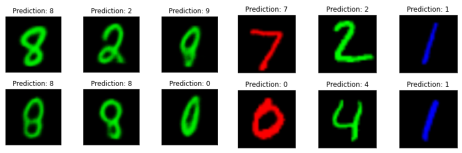

Figure 5: Downstream prediction of a classifier on color MNIST.

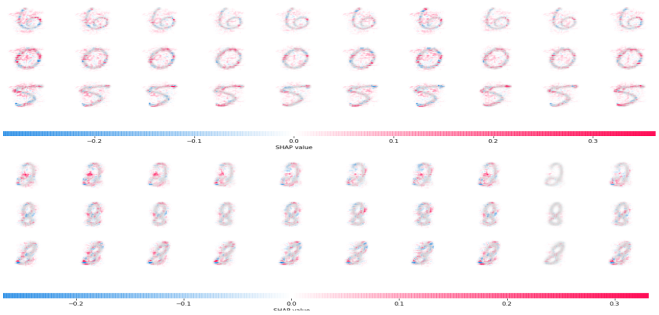

Figure 6: SHAP analysis on color MNIST

7.3 *SHAP on CMIST prediction*

For the image (color MNIST), the most important areas of digits are concentrated on the shape in generated data (Figure 6 on the top), whereas in real data, it is distributed over the entire area including the background (Figure 6 on the bottom).

Architecture	batch size	optimizer	learning rate	epochs	timesteps (T)	number of critic
FLD-WGAN-GP	2048	Adam	2e-3	20000	×	5
FLD-LSGAN	2048	Adam	2e-3	20000	×	×
FLD-DM	2048	Adam	1e-4	5000	1000	×

Table 1: Hyperparameters for our models

Architecture	Layer dimension
Generator	$64 \to 128$ $128 \to 256$ $256 \to 128$ $128 \to 64$ $64 \to 8$
Discriminator	$8 \to 128$ $128 \to 64$ $64 \to 32$ $32 \to 1$
Diffusion	$8 \to 256$ $256 \to 256$ $256 \to 8$

Table 2: Layer Dimension for the Generator, Discriminator

Model	(i)	(ii)	(iii)	(iv)	Goal
VFAE[9]	✓	×	×	↑	synthetic data
DM[6]	×	×	✓	↑	synthetic data
FairGAN [15]	✓	×	×	↑	fair synthetic data
DECAF [13]	✓	×	✓	↑	fair synthetic data
Fair latent deep (ours) generative models	✓	✓	✓	↓	fair synthetic data

Table 3: Overview of related works. The key areas of interest are (i) provide fairness, (ii) syntax-agnostic generation, (iii) fairness optimization is separated from DGP, and (iv) computational overhead (↑ - high, ↓ - low).

References

[1] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena, 'Discriminator rejection sampling', *arXiv preprint arXiv:1810.06758*, (2018).

[2] CB Bell, 'Mutual information and maximal correlation as measures of dependence', *The Annals of Mathematical Statistics*, 587–595, (1962).

[3] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, (2015).

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial networks', *Communications of the ACM*, **63**(11), 139–144, (2020).

[5] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, 'Improved training of wasserstein gans', *Advances in neural information processing systems*, **30**, (2017).

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel, 'Denoising diffusion probabilistic models', *Advances in Neural Information Processing Systems*, **33**, 6840–6851, (2020).

[7] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo, 'Learning debiased representation via disentangled feature augmentation', *Advances in Neural Information Processing Systems*, **34**, 25123–25133, (2021).

[8] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong, 'Fair representation learning: An alternative to mutual information', in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, (2022).

[9] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel, 'The variational fair autoencoder', in *ICLR*, (2016).

[10] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, 'Least squares generative adversarial networks', in *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, (2017).

[11] Junhyun Nam, Sangwoo Mo, Jaeho Lee, and Jinwoo Shin, 'Breaking the spurious causality of conditional generation via fairness intervention with corrective sampling', *arXiv preprint arXiv:2212.02090*, (2022).

[12] Alfréd Rényi, 'On measures of dependence', *Acta mathematica hungarica*, **10**(3-4), 441–451, (1959).

[13] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar, 'Decaf: Generating fair synthetic data using causally-aware generative networks', *Advances in Neural Information Processing Systems*, **34**, 22221–22233, (2021).

[14] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez, 'Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, (2019).

[15] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu, 'Fairgan: Fairness-aware generative adversarial networks', in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, (2018).