IEEE East Asian School of Information Theory
KAIST, August 3 ∼ 6, 2021

August 4, 2021
Changho Suh (chsuh@kaist.ac.kr)

Lecture 5: A fair generative model using total variation distance

Recap

Last time, we investigated the hinge-loss-based GAN [3] which takes the following optimization:

$$\begin{aligned}&\text{Discriminator: } \max_{D} \mathbb{E}_{\mathbb{P}_{\sf real}}\left[\min(0,-1+D(X))\right]+\mathbb{E}_{\mathbb{P}_G}\left[\min(0,-1-D(\tilde{X}))\right];\\ &\text{Generator: } \min_{G} \mathbb{E}_{\mathbb{P}_{\sf real}}\left[D(X)\right]-\mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right].\end{aligned} \tag{1}$$

We then found there is a theoretical insight behind the choice of the hinge loss function for Discriminator and the linear loss function for Generator. The insight was well justified via the interesting connection to a well-known divergence measure, Total Variation Distance (TVD). Plugging the maximizer D^* into the objective function in Generator's optimization, we found:

$$\min_{G} \mathbb{E}_{\mathbb{P}_{\sf real}}\left[D^*(X)\right]-\mathbb{E}_{\mathbb{P}_G}\left[D^*(\tilde{X})\right]=\min_{G} 2\cdot \mathsf{TV}(\mathbb{P}_{\sf real},\mathbb{P}_G). \tag{2}$$

It turns out this connection gives insights into developing a fair generative model. This forms the content of this lecture.

Outline of Lecture 5

In Lecture 5, we will explore a TVD-based fair generative model. This lecture consists of four parts. First we will introduce a TVD-based regularization term that promotes *fair representation* of this tutorial's focus. We will then formulate a corresponding optimization problem. Next we will translate it into an implementable optimization that bears similarity to (1) in structure. Finally, we will discuss experimental results.

A regularization term for fair representation

Observe the TVD-based optimization for Generator in the hinge-loss-based GAN:

$$\min_{G} \mathsf{TV}(\mathbb{P}_{\sf real},\mathbb{P}_G). \tag{3}$$

A question of interest is: How to incorporate a fairness constraint that respects fair representation? To gain insights, let us first ponder on what we want for fair representation. Suppose that $\mathbb{P}_{\sf ref}$ indicates a reference distribution that well respects fair representation. Then, we can naturally come up with a constraint on $\mathbb{P}_G$. That is,

$$\mathbb{P}_G \approx \mathbb{P}_{\sf ref}. \tag{4}$$

However, there is an issue in satisfying the constraint (4). The issue is that the given real data does not necessarily satisfy $\mathbb{P}_{\sf ref}$. Actually what we are targeting are the challenging situations in which $\mathbb{P}_{\sf real}$ is far from $\mathbb{P}_{\sf ref}$ and hence fair representation is not guaranteed with a naive generative model relying solely upon the given real data.

One natural way to address the issue is to employ a new yet small reference dataset that respects $\mathbb{P}_{\sf ref}$. Here what I mean by the degree of small is less than 10% reference dataset relative to the original real dataset. This is because around 10% hand-crafted datasets may be available in practice. This then naturally motivates us to consider $\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$ as a regularization term, since minimizing $\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$ promotes fair representation.

TVD-based optimization for a fair generative model [1]

Inspired by this, Um and Suh [1] came up with the following optimization for fair representation:

$$\min_G (1-\lambda) \cdot \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G) + \lambda \cdot \mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G) \tag{5}$$

where $0 \leq \lambda \leq 1$ denotes a regularization factor that acts as a fairness tuning knob.

An equivalent form based on hinge loss

How to solve the optimization (5)? To gain insights, remember what we did for the hinge-loss-based GAN. Recall that $\mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G)$ was a consequence of evaluating Generator's objective function at the optimal D^*, which was derived from the following Discriminator:

$$\max_D \mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[\min(0, -1 + D(X))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D(\tilde{X}))\right]. \tag{6}$$

So we can now guess the same thing yet w.r.t. $\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$. We conjecture that $\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$ is a consequence of evaluating Generator's objective function for another optimal discriminator, say D^*_{ref}. The optimization problem for D^*_{ref} would be of exactly the same structure yet employing different entities $(D_{\mathsf{ref}}, \mathbb{P}_{\mathsf{ref}}, X_{\mathsf{ref}})$:

$$\max_{D_{\mathsf{ref}}} \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[\min(0, -1 + D_{\mathsf{ref}}(X_{\mathsf{ref}}))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D_{\mathsf{ref}}(\tilde{X}))\right]. \tag{7}$$

As you may easily guess, the conjecture is true. In other words, the optimization (5) is equivalent to solving the following three-level optimization:

$$\begin{aligned}
&\text{Discriminator 1}: \max_D \mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[\min(0, -1 + D(X))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D(\tilde{X}))\right];\\
&\text{Discriminator 2}: \max_{D_{\mathsf{ref}}} \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[\min(0, -1 + D_{\mathsf{ref}}(X_{\mathsf{ref}}))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D_{\mathsf{ref}}(\tilde{X}))\right];\\
&\text{Generator}: \min_G (1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]\right\} + \lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[D_{\mathsf{ref}}(X_{\mathsf{ref}})\right] - \mathbb{E}_{\mathbb{P}_G}\left[D_{\mathsf{ref}}(\tilde{X})\right]\right\}.
\end{aligned} \tag{8}$$

Here Generator's objective function is simply the λ-weighted summation of two individual objectives, each corresponding to each discriminator.

Proof of equivalence

What we need to show is that plugging the optimal $(D^*, D^*_{\mathsf{ref}})$ into Generator's objective function yields:

$$\begin{aligned}
&(1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[D^*(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D^*(\tilde{X})\right]\right\} + \lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[D^*_{\mathsf{ref}}(X_{\mathsf{ref}})\right] - \mathbb{E}_{\mathbb{P}_G}\left[D^*_{\mathsf{ref}}(\tilde{X})\right]\right\}\\
=&2(1-\lambda)\mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G) + 2\lambda\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G).
\end{aligned} \tag{9}$$

The proof is the same as before. Using the technique based on the lemma introduced earlier, one can first show:

$$\begin{aligned}
D^*(x) &= \mathsf{sign}(\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x));\\
D^*_{\mathsf{ref}}(x) &= \mathsf{sign}(\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x)).
\end{aligned} \tag{10}$$

Plugging these into Generator's objective function, we get:

$$
\begin{aligned}
&(1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}[D^*(X)]-\mathbb{E}_{\mathbb{P}_G}\left[D^*(\tilde{X})\right]\right\}+\lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D^*_{\mathsf{ref}}(X_{\mathsf{ref}})]-\mathbb{E}_{\mathbb{P}_G}\left[D^*_{\mathsf{ref}}(\tilde{X})\right]\right\}\\
=&(1-\lambda)\sum_{x\in\mathcal{X}\cup\tilde{\mathcal{X}}}(\mathbb{P}_{\mathsf{real}}(x)-\mathbb{P}_G(x))D^*(x)+\lambda\sum_{x\in\mathcal{X}_{\mathsf{ref}}\cup\tilde{\mathcal{X}}}(\mathbb{P}_{\mathsf{ref}}(x)-\mathbb{P}_G(x))D^*_{\mathsf{ref}}(x)\\
=&(1-\lambda)\sum_{x\in\mathcal{X}\cup\tilde{\mathcal{X}}}|\mathbb{P}_{\mathsf{real}}(x)-\mathbb{P}_G(x)|+\lambda\sum_{x\in\mathcal{X}_{\mathsf{ref}}\cup\tilde{\mathcal{X}}}|\mathbb{P}_{\mathsf{ref}}(x)-\mathbb{P}_G(x)|\\
=&2(1-\lambda)\mathsf{TV}(\mathbb{P}_{\mathsf{real}},\mathbb{P}_G)+2\lambda\mathsf{TV}(\mathbb{P}_{\mathsf{ref}},\mathbb{P}_G)
\end{aligned}
\tag{11}
$$

where $\mathcal{X}_{\mathsf{ref}}$ indicates the set of the samples in the reference dataset.

Architecture

The architecture of the hinge-loss-based optimization (8) is illustrated in Fig. 1. We have one generator and two discriminators: one for intending to synthesize high-quality realistic samples; and the other for promoting fair representation with the help of an additional reference dataset. So one can interpret this as a *three-player game*. We can indeed see three-way battles.

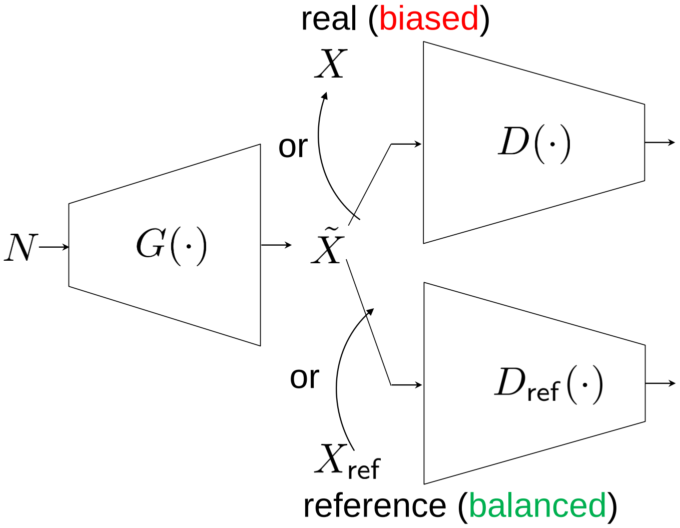

Figure 1: Architecture of the TVD-based fair generative model: A three-player game. An input to $D(\cdot)$ is either X (a real sample) or $\tilde{X}$ (a generated sample). An input to $D_{\mathsf{ref}}(\cdot)$ is either $\tilde{X}$ (a generated sample) or X_{ref} (a reference sample).

The first is a well-known battle between the generator G and the 1st discriminator D. Remember $D^*(x) = \mathsf{sign}(\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x))$. So one can interpret D^* as the strength of distinguishing real (potentially biased) samples against generated samples. On the other hand, the generator intends to fool D, thus promoting realistic samples. The second battle is in between the generator and the 2nd discriminator D_{ref}. The same interpretation can be made from $D^*_{\mathsf{ref}}(x) = \mathsf{sign}(\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x))$ (the ability to distinguish balanced reference samples against the generated samples). This way, the generator G is encouraged to produce balanced (yet possibly unrealistic[1] samples, thus pitting the 1st discriminator against the 2nd discriminator indirectly. The last battle is in between the 1st and 2nd discriminators. This tension is directly controlled by the fairness tuning knob λ. It turns out the three-way tradeoff relationships established

[1] This is because we employ a much smaller-sized reference dataset relative to real dataset.

via the TVD-based framework (5) are greatly balanced, thus achieving significant performances both in fairness and sample quality. This will be empirically demonstrated in the next sections.

Experiment setting

We provide experimental results for a benchmark real dataset CelebA [2] to demonstrate that the TVD-based fair generative model can indeed produce realistic and balanced fake samples. Here is detail of the experiment setting. The real dataset is constructed to have $9:1$ ratio (female vs. male) samples where $m_{\textsf{real}} = 67,507$. We take the balanced samples for the reference dataset ($1:1$ ratio, i.e., $\mathbb{P}_{\textsf{ref}}(z) \sim \textsf{Uniform}$). We consider a realistic scenario in which the size of the additional reference dataset is more or less small, around 10% relative to the real dataset, i.e., $m_{\textsf{ref}} \approx 0.1 m_{\textsf{real}}$.

Figure 2: Real data samples of CelebA.

Performance measures

We consider two performance measures: one for the quality of generated samples; the other for quantifying the degree of fair representation. As for the quality of samples, we employ a well-known score, named Fréchet Inception Distance (FID) [5]. It is defined as the 2nd order Wasserstein distance between the Gaussianized distributions of real and generated samples:

$$\textsf{FID} := W^2(\mathcal{N}(\mu_{\textsf{real}}, \Sigma_{\textsf{real}}), \mathcal{N}(\mu_G, \Sigma_G)) \tag{12}$$

where $(\mu_{\textsf{real}}, \Sigma_{\textsf{real}})$ denote the empirical mean and covariance of real samples, respectively. Similarly for (μ_G, Σ_G). As a measure for fair representation, we consider Fairness Discrepancy (FD) mentioned in Lecture 1 [6]:

$$\textsf{FD} := \sqrt{\sum_{z \in \mathcal{Z}} (\mathbb{P}_G(z) - \mathbb{P}_{\textsf{ref}}(z))^2}. \tag{13}$$

Experimental results

All the simulation results are the ones averaged over five trials with distinct random seeds. Fig. 3 provides FID-vs-FD tradeoff performances for (non-fair) hinge-loss-based GAN and the TVD-based fair generative model. The specific architecture that we employ is [4]. Notice that the TVD-based fair generative model significantly boots up the fairness performance (i.e., reduces FD) with relatively smaller performance degradation in sample quality ("FID"). The lower, the better for all the measures.

	FID (quality)	**FD (fairness)**
Non-*fair* model (Hinge GAN)	8.76 ± 0.196	0.539 ± 0.002
TVD-based *fair* gen. model	14.13 ± 0.343	0.0431 ± 0.0097

Figure 3: FID (quality) vs FD (fairness) tradeoff performances. Here we set a hyperparameter λ as 0.925.

Fig. 4 visualizes generated samples on CelebA. Faces above the yellow line are female samples, while the rest are male samples. In addition to fair representation, the considered approach produces realistic sample images, reflected in a lower FID, around 11.49.

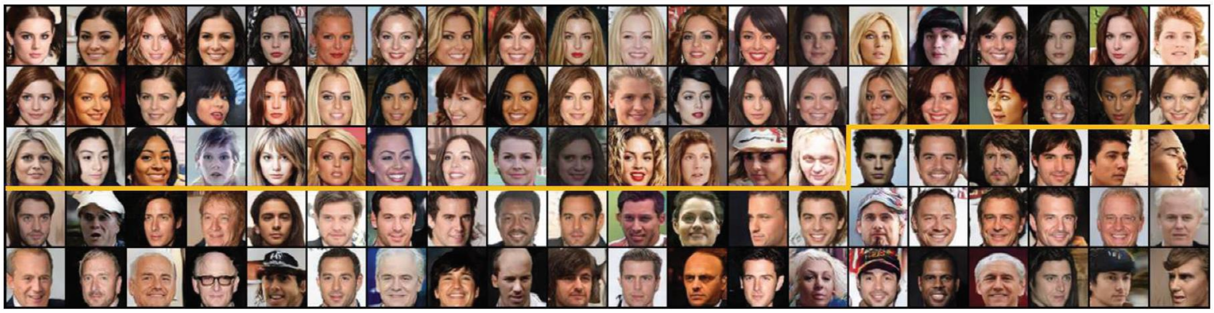

Figure 4: Fake samples generated by the TVD-based fair generative model. Female : male = 54 : 46.

Look ahead

Now we are done for exploring the two focused contexts: (i) fair classifiers; and (ii) fair generative models. In the next (last) lecture, we will conclude the tutorial by saying a few words about two things: (i) fair classifiers being *robust* against data poisoning; and (ii) other contexts beyond the two.

References

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

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

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

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

[5] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[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.