Fair machine learning

Lecture 4

Changho Suh
EE, KAIST

Aug. 4, 2021

A generative model: A hinge-loss-based GAN

Reading: TN4

Recap: Fair classifiers

1. Explore fairness measures in fair classifiers:
   DDP and DEO

2. Study an MI-based fair classifier which offers a good tradeoff while suffering from training instability.

3. Investigate another fair classifier based on KDE, which addresses the training instability issue.

Outline of Lectures 4/5/6

Move onto a different context: **Fair generative model.**

1. Study a generative model which forms the basis of a fair generative model to be explored in depth:

   A hinge-loss-based GAN

2. Make a connection with a well-known statistical measure prevalent in information theory:

   Total Variation Distance (TVD)

3. Investigate a TVD-based fair generative model.

4. Discuss a couple of other relevant issues.

Focus of Lecture 4

Move onto a different context: **Fair generative model.**

1. Study a generative model which forms the basis of a fair generative model to be explored in depth:

   A hinge-loss-based GAN

2. Make a connection with a well-known statistical measure prevalent in information theory:

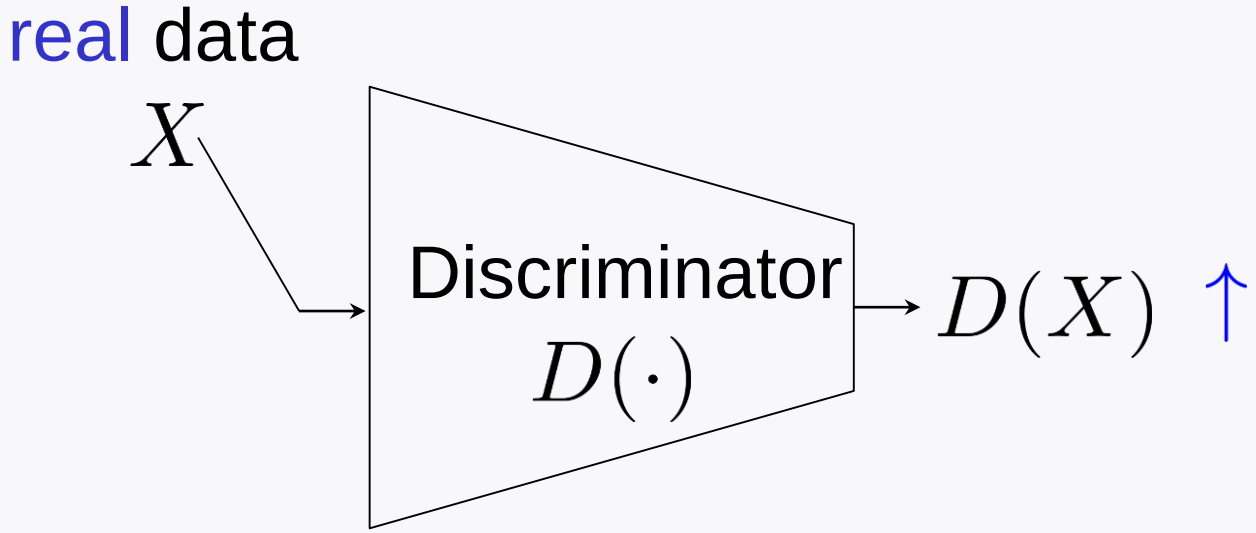
   Total Variation Distance (TVD)

3. Investigate a TVD-based fair generative model.

4. Discuss a couple of other relevant issues.

Generative Adversarial Networks

Goodfellow et al.
NeurIPS14

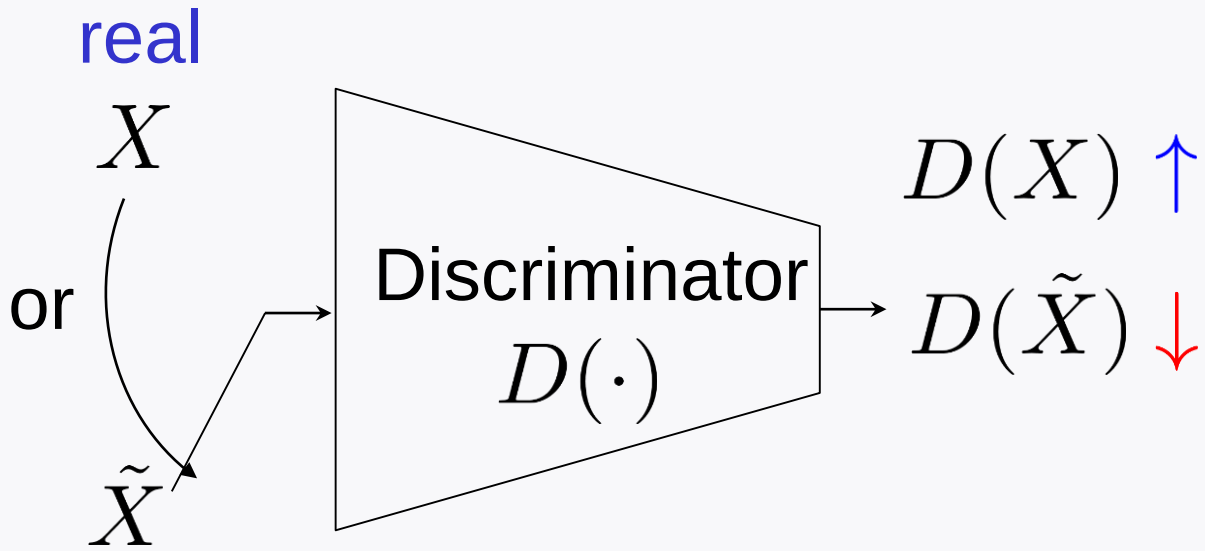


Role: Discriminate real from generated fake samples

Intend to yield a large $D(\cdot)$ if the input is real data;

Generative Adversarial Networks

Goodfellow et al.
NeurIPS14

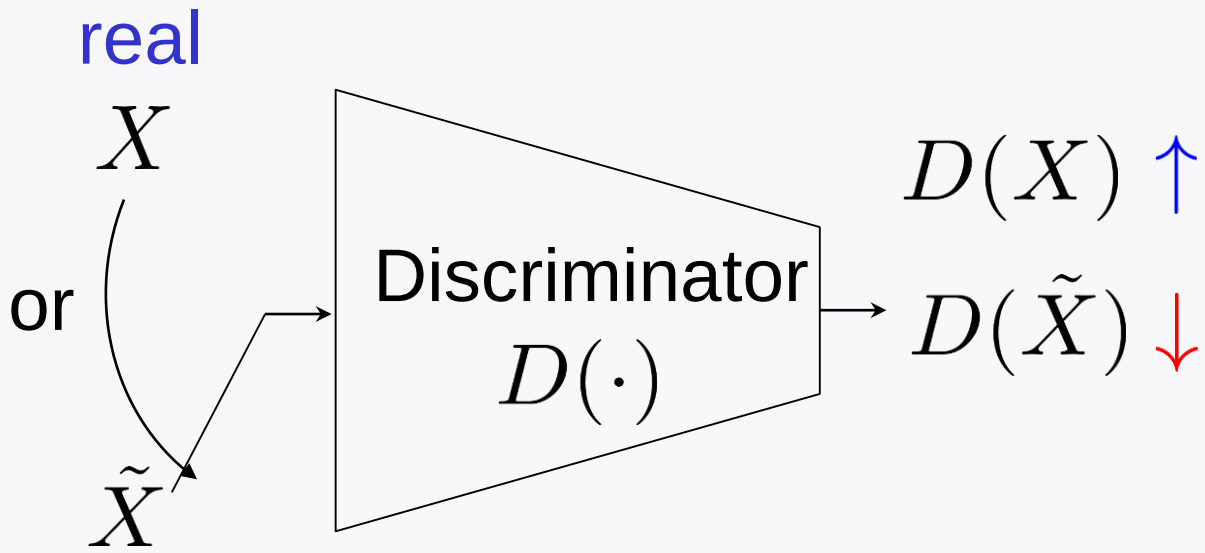


Role: Discriminate real from generated fake samples

Intend to yield a large $D(\cdot)$ if the input is real data;

a small $D(\cdot)$ for generated data.

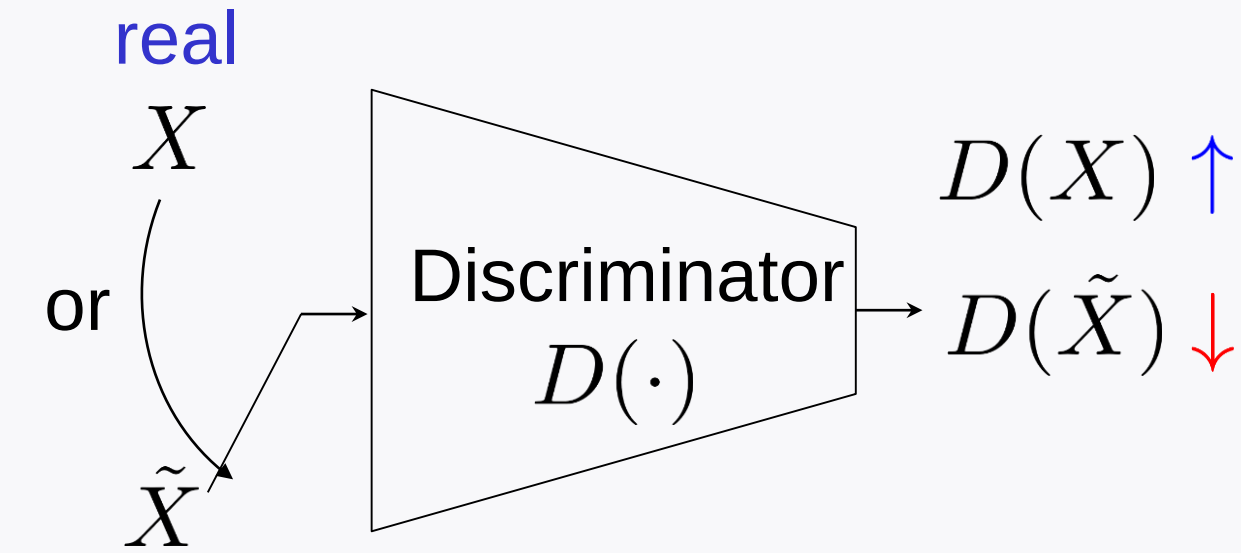
A reasonable interpretation on $D(\cdot)$



Probability of the input being real:

$$D(\cdot\ \ ) = \mathbb{P}((\cdot) = \text{real})$$

A reasonable interpretation on $D(\cdot)$

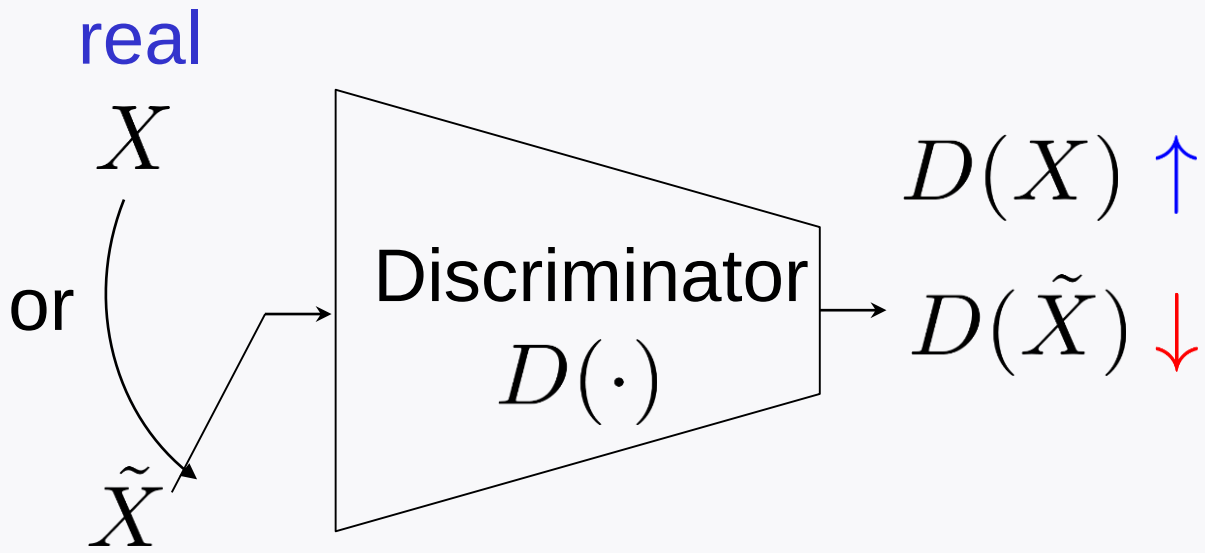


Probability of the input being real:

$$\uparrow \quad D(X) = \mathbb{P}(X = \text{real}) = 1$$

$$\downarrow \quad D(\tilde{X}) = \mathbb{P}(\tilde{X} = \text{real}) = 0$$

Goal



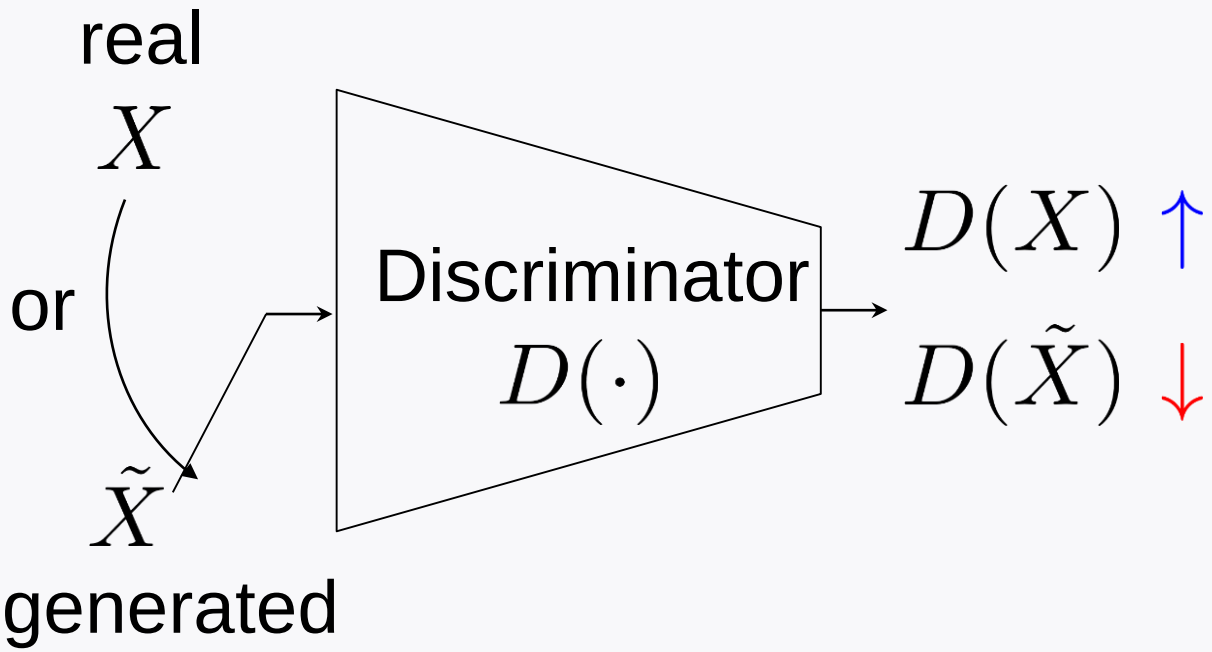
Maximize: $D(X)$ & $1 - D(\tilde{X})$

A natural optimization:

$$\max_D \mathbb{E}_{\mathbb{P}_{\text{real}}}[D(X)] + \mathbb{E}_{\mathbb{P}_G}\left[1 - D(\tilde{X})\right]$$

Log loss

Goodfellow et al. NeurIPS14



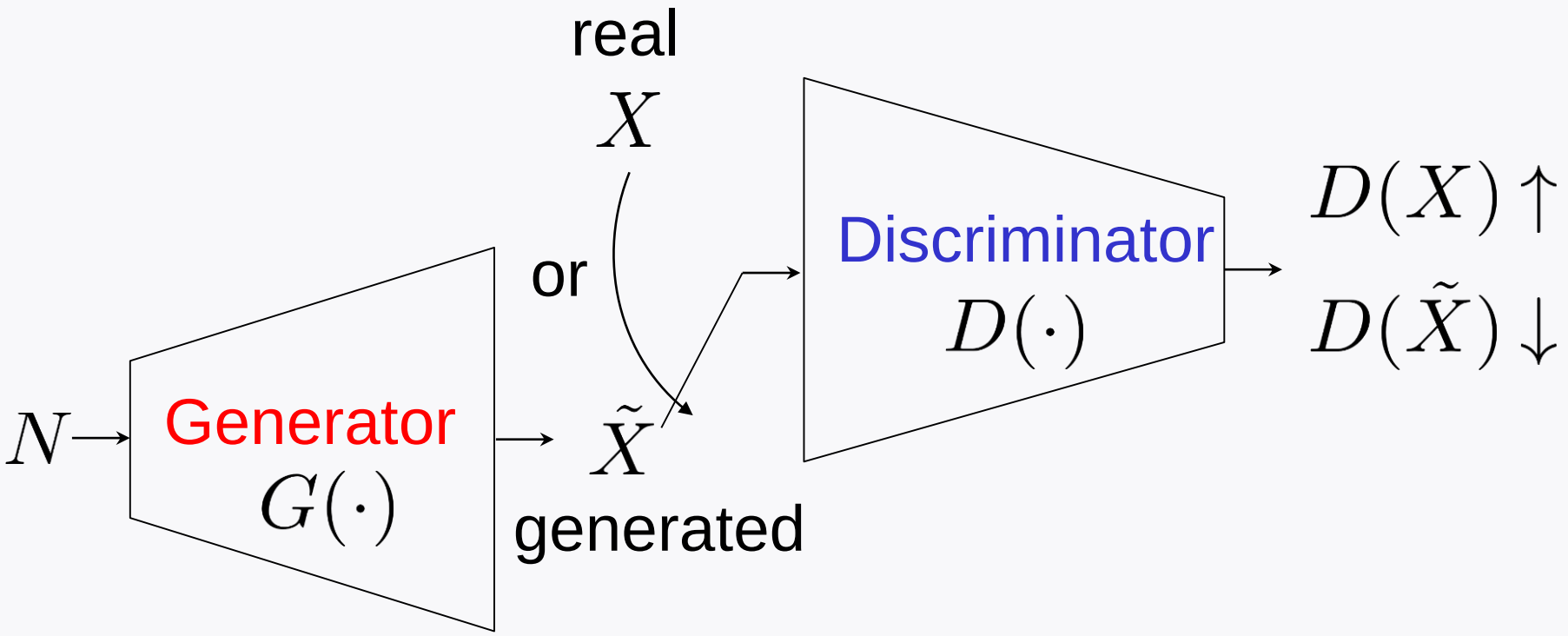
Goodfellow employed log loss instead:

$$\max_D \mathbb{E}_{\mathbb{P}_{\text{real}}} [\log D(X)] + \mathbb{E}_{\mathbb{P}_G} \left[\log(1 - D(\tilde{X}))\right]$$

Goodfellow mentioned: It is inspired by [Gutmann et al. AISTATS10] in his NeurIPS16's tutorial.

A two-player game

Goodfellow et al. NeurIPS14



Discriminator: $\max_D \mathbb{E}_{\mathbb{P}_{\text{real}}} [\log D(X)] + \mathbb{E}_{\mathbb{P}_G} \left[\log(1 - D(\tilde{X}))\right]$

Generator: $\min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} [\log D(X)] + \mathbb{E}_{\mathbb{P}_G} \left[\log(1 - D(\tilde{X}))\right]$

A variant of GAN

Lim et al. arxiv17

GAN:

D: $\max_D \mathbb{E}_{\mathbb{P}_{\text{real}}} [\log D(X)] + \mathbb{E}_{\mathbb{P}_G} \left[\log(1 - D(\tilde{X}))\right]$

G: $\min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} [\log D(X)] + \mathbb{E}_{\mathbb{P}_G} \left[\log(1 - D(\tilde{X}))\right]$

Lim et al. employed different loss functions:

hinge loss for D; linear loss for G.

Hinge-loss-based GAN:

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

G: $\min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} [D(X)] - \mathbb{E}_{\mathbb{P}_G} \left[D(\tilde{X})\right]$

Hinge-loss-based GAN

Lim et al. arxiv17

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

$$G: \min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} [D(X)] - \mathbb{E}_{\mathbb{P}_G} \left[D(\tilde{X})\right]$$

Turns out: Its advanced version offers the state-of-the-art performance in many of the real datasets.

Fake samples generated by its advanced version (CelebA)

Hinge-loss-based GAN

Lim et al. arxiv17

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

$$G: \min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} \left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G} \left[D(\tilde{X})\right]$$

Looks like a heuristic choice on the loss functions.

But it turns out this yields an interesting connection w/:

Total Variation Distance (TVD)

Also the connection gives insights into developing a fair generative model.

Connection with TVD

Tan et al. Stat19

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

$$G: \min_G \mathbb{E}_{\mathbb{P}_{\text{real}}}\left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]$$

One way to solve the optimization:

First solve D^*

Then plug D^* into Generator's objective to solve G^*

Connection with TVD

Tan et al. Stat19

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

$$G: \min_G \mathbb{E}_{\mathbb{P}_{\text{real}}}\left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]$$

Theorem (Tan et al.):

$$D^*(x) = \text{sign}(\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)) \qquad \forall x$$

$$G^* = \arg\min_G \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} |\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)|$$

set of real samples ↗ ↖ set of generated samples

$$\mathbb{P}_{\text{real}}(x) = 0 \quad \text{if } x \in \tilde{\mathcal{X}} \setminus \mathcal{X}$$

$$\mathbb{P}_G(x) = 0 \quad \text{if } x \in \mathcal{X} \setminus \tilde{\mathcal{X}}$$

Connection with TVD

Tan et al. Stat19

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

$$G: \min_G \mathbb{E}_{\mathbb{P}_{\text{real}}}\left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]$$

Theorem (Tan et al.):

$$D^*(x) = \text{sign}(\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)) \qquad \forall x$$

$$G^* = \arg\min_G \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} |\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)|$$

$$=: 2 \cdot \text{TV}(\mathbb{P}_{\text{real}}, \mathbb{P}_G)$$

$$\mathbb{P}_{\text{real}}(x) = 0 \quad \text{if } x \in \tilde{\mathcal{X}} \setminus \mathcal{X}$$

$$\mathbb{P}_G(x) = 0 \quad \text{if } x \in \mathcal{X} \setminus \tilde{\mathcal{X}}$$

Proof: $D^*(x) = \text{sign}(\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)) \qquad \forall x$

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

$$\sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} \mathbb{P}_{\text{real}}(x) \min(0, -1 + D(x)) + \mathbb{P}_G(x) \min(0, -1 - D(x))$$

Using Lemma, we get:

$$D^*(x) = \text{sign}(\mathbb{P}_{\text{real}}(x) - \mathbb{P}_G(x)) \quad \forall x$$

Lemma:

$$f(t) = \alpha \cdot \min\{0, -1 + t\} + \beta \cdot \min\{0, -1 - t\}$$

For given $\alpha, \beta \geq 0$: $t^* = \arg\max_t f(t) = \text{sign}(\alpha - \beta)$

Proof: $G^* = \arg\min_G \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G)$

$$D^*(x) = \mathsf{sign}(\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x)) \qquad \forall x$$

$$\mathsf{G}: \quad \min_G \mathbb{E}_{\mathbb{P}_{\mathsf{real}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]$$

$$\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}[D^*(X)] - \mathbb{E}_{\mathbb{P}_G}\left[D^*(\tilde{X})\right]$$

$$= \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} \mathbb{P}_{\mathsf{real}}(x) D^*(x) - \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} \mathbb{P}_G(x) D^*(x)$$

$$= \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} (\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x)) D^*(x)$$

$$= \sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} |\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x)| = 2 \cdot \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G)$$

Look ahead

Investigate a TVD-based fair generative model.

Reference

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems 27 (NeurIPS)*, 2014.

[2] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *AISTATS* 2010.

[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] Z. Tan, Y. Song, and Z. Ou. Calibrated adversarial algorithms for generative modelling. *Stat*, 2019.

[6] L. Tierney. Introduction to general state-space markov chain theory. *Markov chain Monte Carlo in practice*, 1996.