2021 IEEE East Asian School of Information Theory

Fair machine learning

Lecture 5

Changho Suh

EE, KAIST

Aug. 4, 2021

A fair generative model using total variation distance

Reading: TN5

Recap: Hinge-loss-based GAN

Discriminator (hinge loss):

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

Generator (linear loss):

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

Made a connection w/ **total variation distance**:

$$\min_G \mathbb{E}_{\mathbb{P}_{\text{real}}} [D^*(X)] - \mathbb{E}_{\mathbb{P}_G} \left[D^*(\tilde{X})\right] = \min_G 2 \cdot \text{TV}(\mathbb{P}_{\text{real}}, \mathbb{P}_G)$$

Claimed: This connection gives insights into a fair generative model.

Outline of Lecture 5

Explore a **TVD**-based fair generative model.

1. Introduce a TVD-based regularization term that promotes **fair sample generation (class-balanced samples)**
2. Formulate a corresponding optimization.
3. Translate it into an **implementable** optimization that employs hinge loss.
4. Discuss experimental results.

A regularization term for fair sample gen.

Generator in hinge-loss-based GAN:

$$\min_{G} \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_{G})$$

What we want for fair sample generation:

$\mathbb{P}_G \approx \mathbb{P}_{\mathsf{ref}} \longleftarrow$ a reference distribution respecting fair sample gen.

An issue arises in satisfying this:

The given real data does not necessarily satisfy $\mathbb{P}_{\mathsf{ref}}$.

Note: Interested scenarios: $\mathbb{P}_{\mathsf{real}}$ biased

A regularization term for fair sample gen.

Generator in hinge-loss-based GAN:

$$\min_{G} \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_{G})$$

What we want for fair sample generation:

$$\mathbb{P}_G \approx \mathbb{P}_{\mathsf{ref}} \longleftarrow$$ a reference distribution respecting fair sample gen.

A natural way to satisfy this:

Introduce a *new* yet *small* reference dataset respecting $\mathbb{P}_{\mathsf{ref}}$.

5-10% relative to original real data

A regularization term for fair sample gen.

Generator in hinge-loss-based GAN:

$$\min_G \mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G)$$

What we want for fair sample generation:

$$\mathbb{P}_G \approx \mathbb{P}_{\mathsf{ref}} \longleftarrow$$ a reference distribution respecting fair sample gen.

A nautral regularization term:

$$\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$$

TVD-based optimization for a fair gen. model

[Um-Suh '21]:

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

Question:

How to solve the optimization?

Observation

[Um-Suh '21]:

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

Remember: $\mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G)$ was a consequence of evaluating Generator's objective at D^*, which was derived from:

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

Observation

[Um-Suh '21]:

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

Guess: $\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$ is a consequence of evaluating Generator's objective at another D^*_{ref}, which is derived from:

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

Equivalence

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

Turns out: Equivalent to

$$\max_D \mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[\min(0, -1 + D(X))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D(\tilde{X}))\right]$$
$$\max_{D_{\mathsf{ref}}} \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[\min(0, -1 + D_{\mathsf{ref}}(X_{\mathsf{ref}}))\right] + \mathbb{E}_{\mathbb{P}_G}\left[\min(0, -1 - D_{\mathsf{ref}}(\tilde{X}))\right]$$
$$\min_G (1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}\left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]\right\} + \lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\left[D_{\mathsf{ref}}(X_{\mathsf{ref}})\right] - \mathbb{E}_{\mathbb{P}_G}\left[D_{\mathsf{ref}}(\tilde{X})\right]\right\}$$

To prove this, need to show:

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

Proof of equivalence

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

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

$$\min_{G} (1 - \lambda) \left\{ \mathbb{E}_{\mathbb{P}_{\text{real}}} \left[D(X)\right] - \mathbb{E}_{\mathbb{P}_G} \left[D(\tilde{X})\right] \right\} + \lambda \left\{ \mathbb{E}_{\mathbb{P}_{\text{ref}}} \left[D_{\text{ref}}(X_{\text{ref}})\right] - \mathbb{E}_{\mathbb{P}_G} \left[D_{\text{ref}}(\tilde{X})\right] \right\}$$

Using the technique based on the lemma introduced earlier, one can show:

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

$$D^*_{\text{ref}}(x) = \text{sign}(\mathbb{P}_{\text{ref}}(x) - \mathbb{P}_G(x)) \qquad \forall x \in \mathcal{X}_{\text{ref}} \cup \tilde{\mathcal{X}}$$

Proof of equivalence

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

$$D^*_{\mathsf{ref}}(x) = \mathsf{sign}(\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x)) \quad \forall x \in \mathcal{X}_{\mathsf{ref}} \cup \tilde{\mathcal{X}}$$

$$\min_G (1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}\left[D(\tilde{X})\right]\right\} + \lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D_{\mathsf{ref}}(X_{\mathsf{ref}})] - \mathbb{E}_{\mathbb{P}_G}\left[D_{\mathsf{ref}}(\tilde{X})\right]\right\}$$

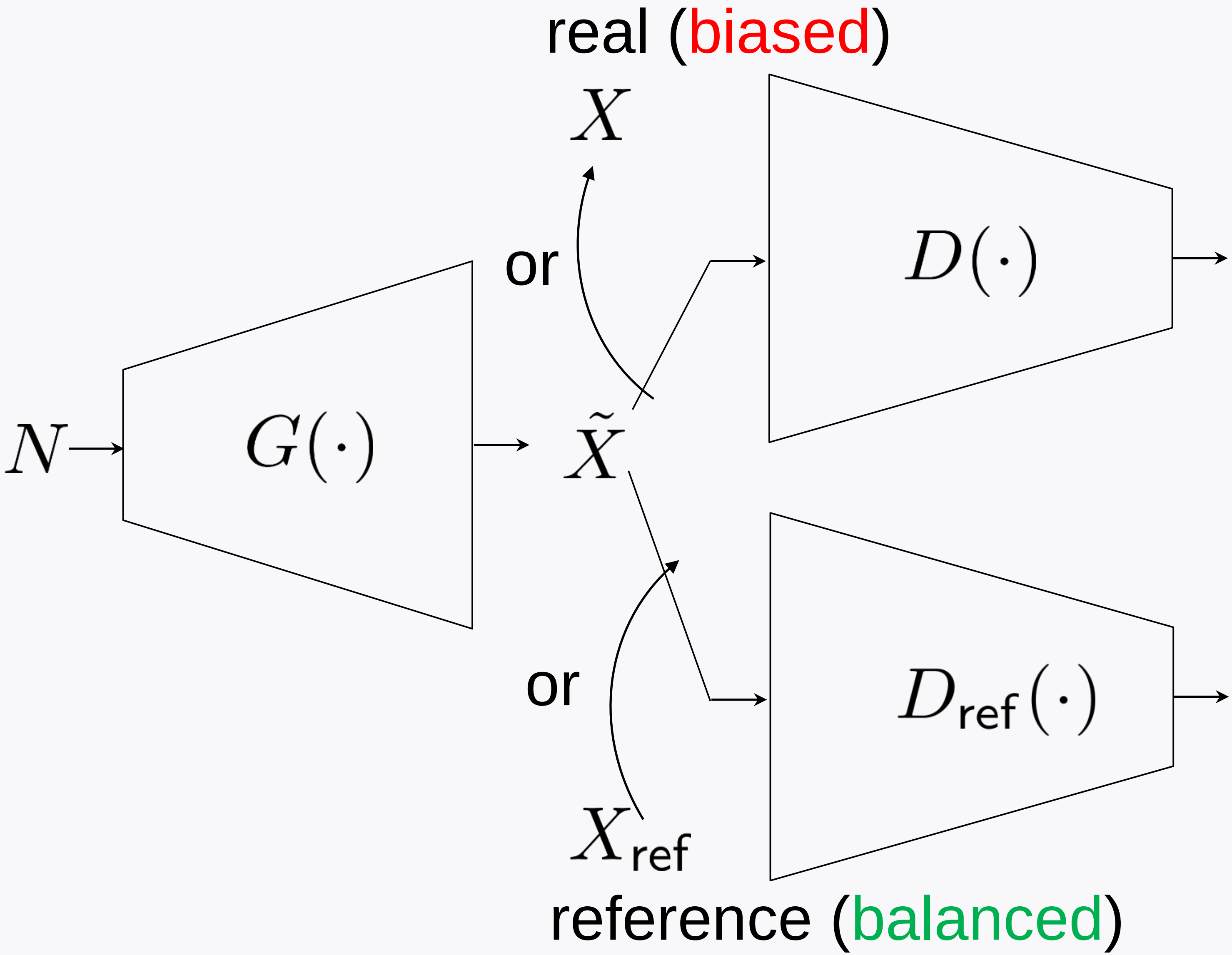
$$(1-\lambda)\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{real}}}[D^*(X)] - \mathbb{E}_{\mathbb{P}_G}\left[D^*(\tilde{X})\right]\right\} + \lambda\left\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D^*_{\mathsf{ref}}(X_{\mathsf{ref}})] - \mathbb{E}_{\mathbb{P}_G}\left[D^*_{\mathsf{ref}}(\tilde{X})\right]\right\}$$

$$= (1-\lambda)\sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} (\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x))D^*(x) + \lambda \sum_{x \in \mathcal{X}_{\mathsf{ref}} \cup \tilde{\mathcal{X}}} (\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x))D^*_{\mathsf{ref}}(x)$$

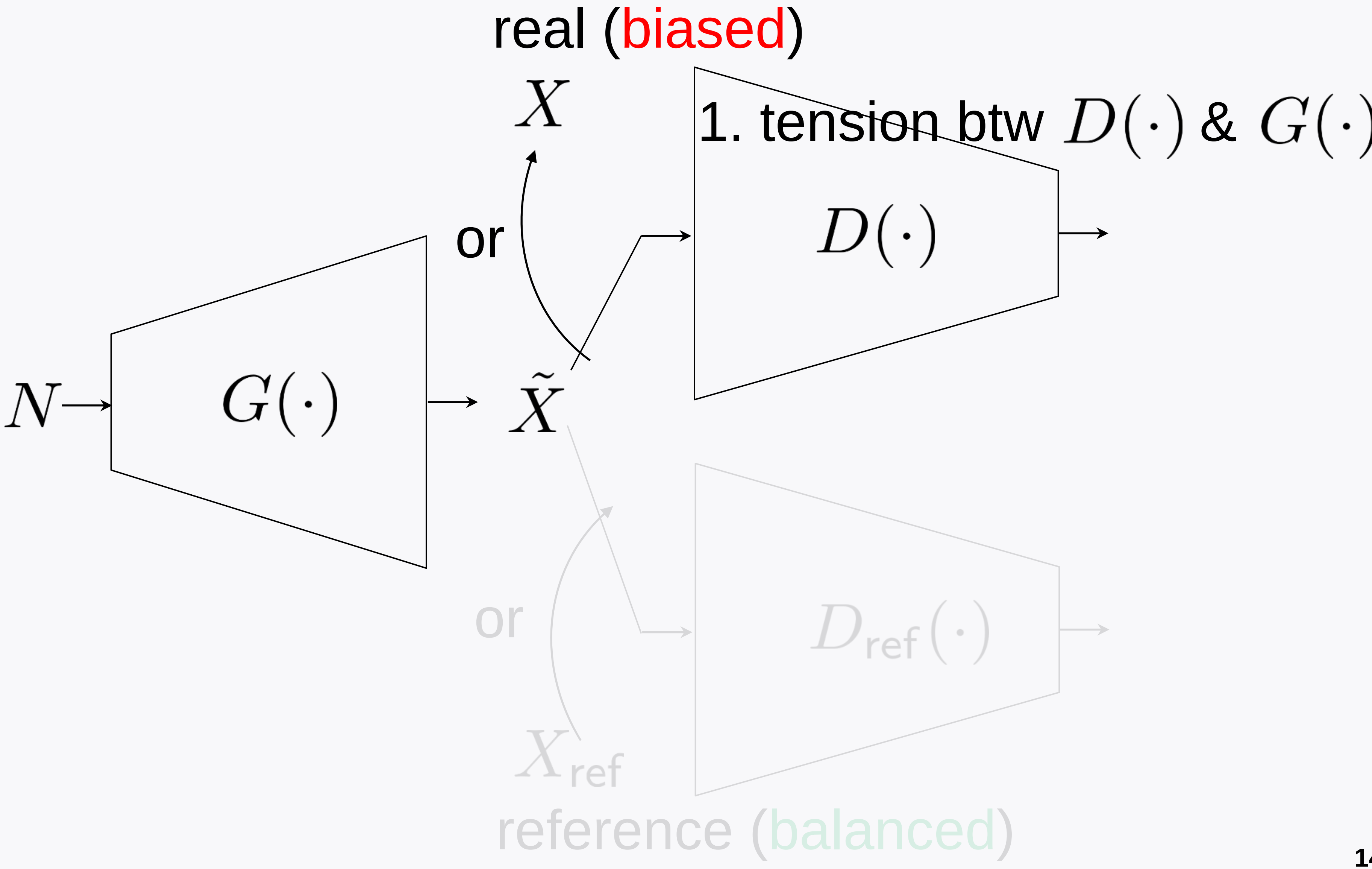
$$= (1-\lambda)\sum_{x \in \mathcal{X} \cup \tilde{\mathcal{X}}} |\mathbb{P}_{\mathsf{real}}(x) - \mathbb{P}_G(x)| + \lambda \sum_{x \in \mathcal{X}_{\mathsf{ref}} \cup \tilde{\mathcal{X}}} |\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x)|$$

$$= 2(1-\lambda)\mathsf{TV}(\mathbb{P}_{\mathsf{real}}, \mathbb{P}_G) + 2\lambda\mathsf{TV}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G)$$

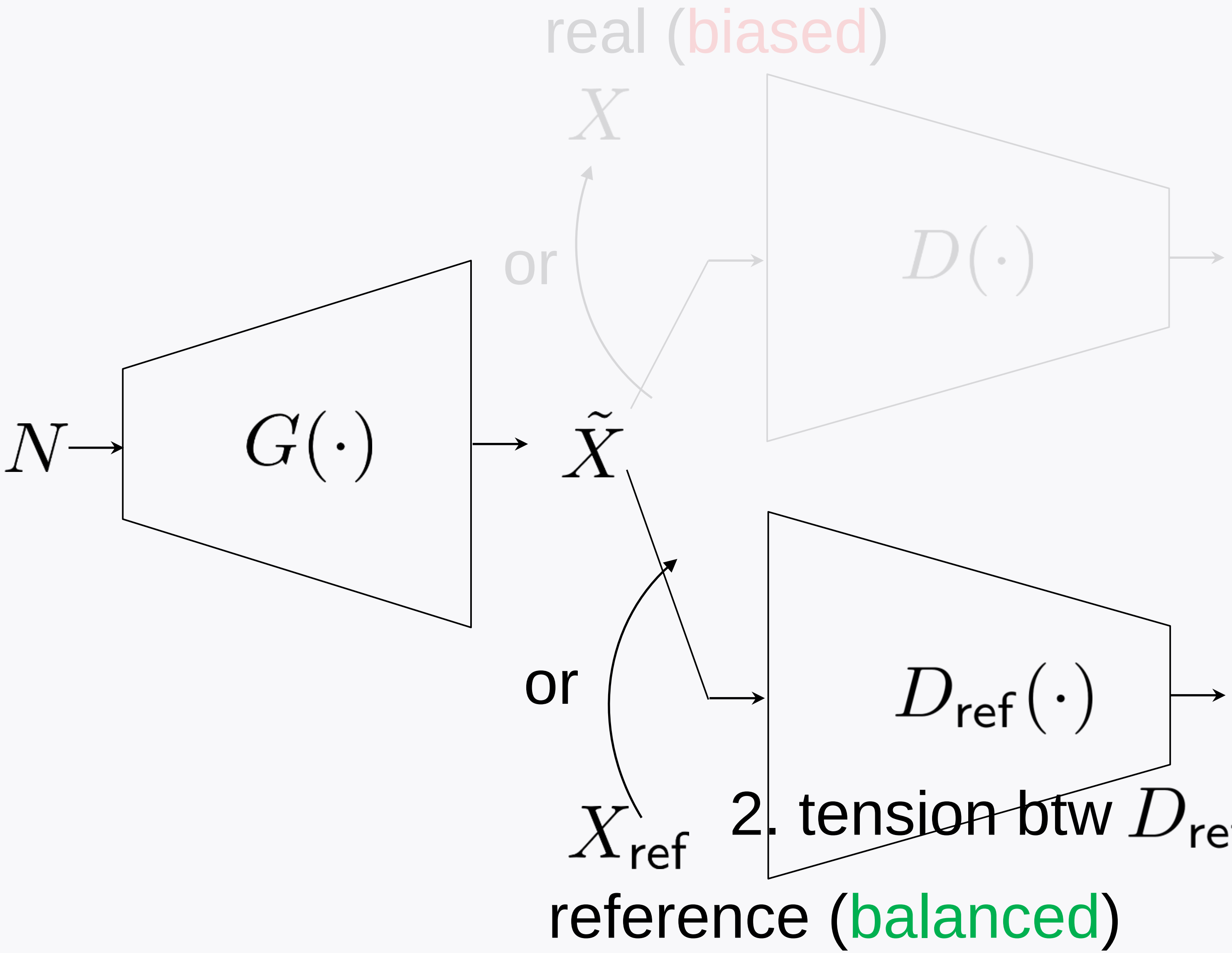
Architecture



Three-way battle

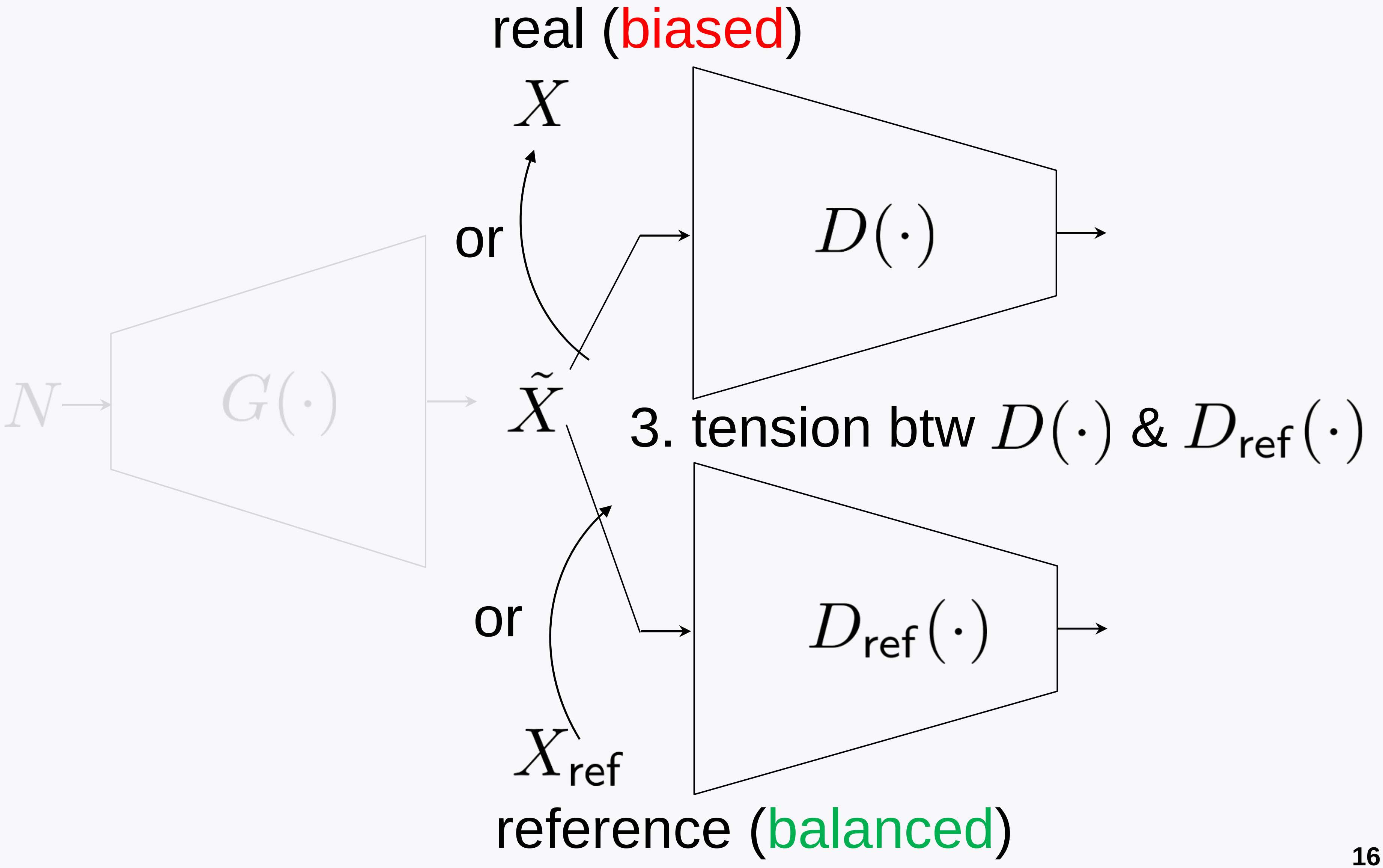


Three-way battle



2. tension btw $D_{\text{ref}}(\cdot)$ & $G(\cdot)$

Three-way battle



Experiments

A benchmark real dataset: **CelebA**

Female:male ~= 90:10

$$m_{\text{real}} = 67,507$$

$$m_{\text{ref}} \approx 0.1 m_{\text{real}}$$

$$\mathbb{P}_{\text{ref}}(z) \sim \text{Uniform}$$

Performance measure

1. A measure for the quality of generated samples

   Fréchet Inception Distance (FID):

$$W^2(\mathcal{N}(\mu_{\text{real}}, \Sigma_{\text{real}}), \mathcal{N}(\mu_G, \Sigma_G))$$

2^{nd} order Wasserstein distance

2. A measure for fair sample generation

   Fairness Discrepancy (FD):

$$\sqrt{\sum_{z \in \mathcal{Z}} (\mathbb{P}_G(z) - \mathbb{P}_{\text{ref}}(z))^2}$$

FID vs FD tradeoff

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

Generated samples

TVD-based fair generative model:

Female:male = 54:46

Look ahead

Discuss a couple of other relevant issues.

Reference

[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. *In Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[3] 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. *In Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[4] 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.