Supplementary Material of Improving the Fairness of the Min-Max Game in GANs Training

Zhaoyu Zhang, Yang Hua, Hui Wang, Seán McLoone
Queen's University Belfast
{zzhang55,Y.Hua,h.wang,s.mcloone}@qub.ac.uk

1. Proofs

Theorem 1. Theoretical analysis of IGGAN.

Case 1: IGGAN with NDA in D_1 and PDA in D_2

Let $\bar{P} \in \mathbb{P}(\chi)$ be any distribution over χ with disjoint support than p_{data}, such that $supp(p_{data}) \cap supp(\bar{P}) = \varnothing$. Let $P^T \in p_{data}$ be any distribution over real data. Let $D_1 : \chi \to \mathbb{R}$ and $D_2 : \chi \to \mathbb{R}$ be the set of discriminators over χ, $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a convex, semi-continuous function such that $f(1) = 0$, f^* be the convex conjugate of f, $f^{'}$ be the derivative of f, G_θ be a distribution with sample space χ, and $G_\theta^T \in G_\theta$ be any distribution over sample space χ. T is one kind of PDA method. Then $\forall \lambda \in (0, 1]$, we have

$$\begin{aligned} &\underset{G_\theta \in P(\chi)}{argmin} \underset{D_1, D_2 : \chi \to \mathbb{R}}{max} L_f(G_\theta, D_1, D_2) \\ &= \underset{G_\theta \in P(\chi)}{argmin} \underset{D_1, D_2 : \chi \to \mathbb{R}}{max} L_f(\lambda G_\theta + (1-\lambda)\bar{P}, D_1, D_2) \\ &= p_{data}, \end{aligned} \tag{S1}$$

where $L_f(G_\theta, D_1, D_2) = E_{x \sim p_{data}}[D_1(x)] - E_{x \sim G_\theta}[f^*(D_1(x))] + E_{x \sim p_{data}}[D_2(T(x))] - E_{x \sim G_\theta}[f^*(D_2(T(x)))]$ is the objective function for IGGAN following NDA-GAN [5] and f-GAN [4]. The optimal discriminators for D_1 and D_2 are different, shown as follows:

$$\begin{aligned} &\underset{D_1 : \chi \to \mathbb{R}}{argmax} L_f(\lambda G_\theta + (1-\lambda)\bar{P}, D_1) \\ &= f^{'}(p_{data}/(\lambda G_\theta + (1-\lambda)\bar{P}). \end{aligned} \tag{S2}$$

$$\underset{D_2 : \chi \to \mathbb{R}}{argmax} L_f(G_\theta, D_2) = f^{'}(P^T / G_\theta^T). \tag{S3}$$

Proof. Let $p(x)$, $\bar{p}(x)$, $p^T(x)$ and $q(x)$ denote the density functions of p_{data}, $\bar{P}$, P^T and G_θ respectively (and P, $\bar{P}$, P^T, Q for the respective distributions). For D_1, following Theorem 1 and Appendix C as in [5], we can obtain the conclusion that we must have $q(x) = p(x)$ for all $x \in \chi$ in D_1. Thus, the generator distribution recovers the data distribution at the Nash equilibrium. For D_2, according to Theorem 1 and section V.B in [6], the generator distribution still recovers the data distribution at the Nash equilibrium.

To sum up, the generator distribution recovers the data distribution at the Nash equilibrium for both D_1 and D_2, which guarantees the convergence of IGGAN.

Moreover, from Lemma 1 in [3], we have that

$$\underset{D : \chi \to \mathbb{R}}{argmax} L_f(Q, D) = f^{'}(p_{data}/Q). \tag{S4}$$

Therefore, by replacing Q with $\lambda G_\theta + (1-\lambda)\bar{P}$ in D_1 and p_{data} with P^T as well as Q with G_θ^T in D_2, we have

$$\begin{aligned} &\underset{D_1 : \chi \to \mathbb{R}}{argmax} L_f(\lambda G_\theta + (1-\lambda)\bar{P}, D_1) \\ &= f^{'}(p_{data}/(\lambda G_\theta + (1-\lambda)\bar{P}), \end{aligned} \tag{S5}$$

$$\underset{D_2 : \chi \to \mathbb{R}}{argmax} L_f(G_\theta, D_2) = f^{'}(P^T / G_\theta^T). \tag{S6}$$

Eq.(S5) and Eq.(S6) show that the optimal discriminators are indeed different for the D_1 and D_2. □

Case 2: IGGAN with different PDAs in D_1 and D_2

Let $P^{T_1}, P^{T_2} \in p_{data}$ be any distribution over real data. Let $D_1 : \chi \to \mathbb{R}$ and $D_2 : \chi \to \mathbb{R}$ be the set of discriminators over χ, $f : \mathbb{R}_{\geq 0} \to \mathbb{R}$ be a convex, semi-continuous function such that $f(1) = 0$, f^* be the convex conjugate of f, $f^{'}$ be the derivative of f, G_θ be a distribution with sample space χ, and $G_\theta^{T_1}, G_\theta^{T_2} \in G_\theta$ be any distribution over sample space χ. T_1 and T_2 are different PDA methods. Then $\forall \lambda \in (0, 1]$, we have

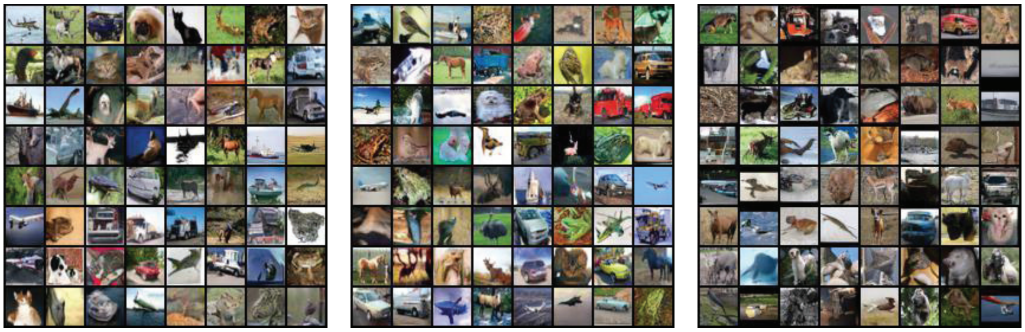
$$\underset{G_\theta \in P(\chi)}{argmin} \underset{D_1, D_2 : \chi \to \mathbb{R}}{max} L_f(G_\theta, D_1, D_2) = p_{data}, \tag{S7}$$

where $L_f(G_\theta, D_1, D_2) = E_{x \sim p_{data}}[D_1(T_1(x)] - E_{x \sim G_\theta}[f^*(D_1(T_1(x))] + E_{x \sim p_{data}}[D_2(T_2(x))] - E_{x \sim G_\theta}[f^*(D_2(T_2(x)))]$ is the objective function for IGGAN following NDA-GAN [5] and f-GAN [4]. The optimal discriminators for D_1 and D_2 are different, shown as follows:

$$\underset{D_1 : \chi \to \mathbb{R}}{argmax} L_f(G_\theta, D_1) = f^{'}(P^{T_1} / G_\theta^{T_1}). \tag{S8}$$

$$\underset{D_2 : \chi \to \mathbb{R}}{argmax} L_f(G_\theta, D_2) = f^{'}(P^{T_2} / G_\theta^{T_2}). \tag{S9}$$

Proof. Let $p(x)$, $p^{T_1}(x)$, $p^{T_2}(x)$ and $q(x)$ denote the density functions of p_{data}, P^{T_1}, P^{T_2} and G_θ respectively (and P,



(a) (b) (c)

Figure S1. Images generated on the CIFAR-10 and STL-10 datasets by IGGAN: (a) Unconditional generation results on CIFAR-10 by IGGAN with Diff-Augment as PDA and Jigsaw as NDA (FID 10.68). (b) Conditional generation results on CIFAR-10 by IGGAN with Diff-Augment as PDA and Cutmix as NDA (FID 8.15). Unconditional generation results on STL-10 by IGGAN with Diff-Augment as PDA and Jigsaw as NDA (FID 21.39). *Best viewed in color.*

P^{T_1}, P^{T_2}, Q for the respective distributions). For both D_1 and D_2, following Theorem 1 and section V.B in [6], we can conclude that the generator distribution recovers the data distribution at the Nash equilibrium.

To conclude, the generator distribution recovers the data distribution at the Nash equilibrium for both D_1 and D_2, which guarantees the convergence of IGGAN.

Moreover, from Lemma 1 in [3], we have that

$$\underset{D:\chi\rightarrow\mathbb{R}}{argmax} L_f(Q, D) = f^{'}(p_{data}/Q). \tag{S10}$$

Therefore, by replacing p_{data} with P^{T_1} and Q with $G_\theta^{T_1}$ in D_1 and p_{data} with P^{T_2} and Q with $G_\theta^{T_2}$ in D_2, we have

$$\underset{D_1:\chi\rightarrow\mathbb{R}}{argmax} L_f(G_\theta, D_1) = f^{'}(P^{T_1}/G_\theta^{T_1}). \tag{S11}$$

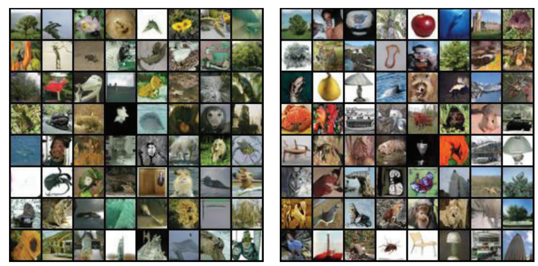
$$\underset{D_2:\chi\rightarrow\mathbb{R}}{argmax} L_f(G_\theta, D_2) = f^{'}(P^{T_2}/G_\theta^{T_2}). \tag{S12}$$

Eq.(S10) and Eq.(S11) show that the optimal discriminators are indeed different for D_1 and D_2. □

The training algorithms of IGGAN (NDA + PDA) and IGGAN (PDA + PDA) are shown in Algorithms 1 and 2, respectively.

2. More Generated Images

According to the main paper, more generated results with IGGAN (BigGAN backbone [1]) on CIFAR-10/STL-10, CIFAR-100 and CelebA are shown in Figures S1, S2 and S3, respectively. More generated results with IGGAN (StyleGAN2 backbone [2]) on FFHQ and LSUNCAT datasets are shown in Figure S4.



(a) (b)

Figure S2. Images generated on the CIFAR-100 dataset by IGGAN: (a) Unconditional generation results on CIFAR-100 by IGGAN with Diff-Augment as PDA and Jigsaw as NDA (FID 16.08). (b) Conditional generation results on CIFAR-100 by IGGAN with Diff-Augment as PDA and Jigsaw as NDA (FID 11.30). *Best viewed in color.*

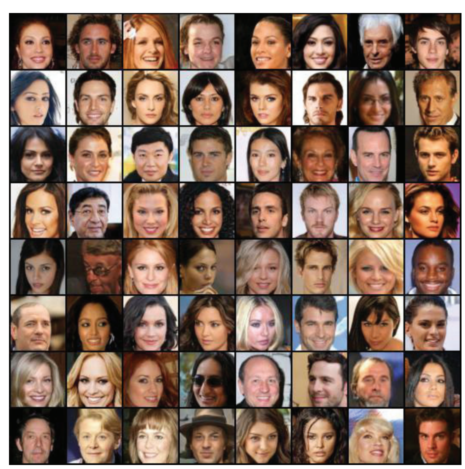
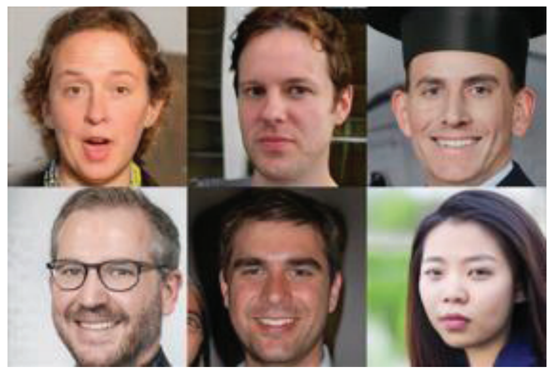


Figure S3. Unconditional images generated on the CelebA dataset by IGGAN (FID 20.63). *Best viewed in color.*

FFHQ-1K FID 20.16

FFHQ-5K FID 9.47

FFHQ-10K FID 7.14



FFHQ-30K FID 4.89

LSUNCAT-1K FID 30.80

LSUNCAT-5K FID 15.85

LSUNCAT-10K FID 11.20

LSUNCAT-30K FID 9.14

Figure S4. Image generated by IGGAN (PDA + PDA) on FFHQ and LSUNCAT datasets. Following Diff-Augment [7], we perform generated results on 30K, 10K, 5K and 1K training samples. *Best viewed in color.*

Algorithm 1 Training algorithm for IGGAN (NDA + PDA).

Require: The number of D_1 iterations n_{D_1}, the number of D_2 iterations n_{D_2}, batchsize $m = 64$, $f_{1_w}(x) = E_{x\sim p_{data}}[D_{1_w}(x)]$ and $f^1_{1_w} = -E_{x\sim G_\theta}[f^*(D_{1_w}(x))]$ determine the objective function of D_{1_w}, $f_{2_w} = E_{x\sim p_{data}}[D_{2_w}(T(x))]$ and $f^2_{2_w} = -E_{x\sim G_\theta}[f^*(D_{2_w}(T(x)))]$ determine the objective function of D_{2_w}, where T is one PDA method. $\bar{P}$ is the distribution of the NDA. w and θ are the parameters of the Ds and G, respectively.

while θ has not converged **do**
  for t=1, ..., n_{D_1} **do**
    Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
    Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
    Update w using SGD by ascending with:
      $\nabla_w \frac{1}{m}\sum_{i=1}^m [f_{1_w}(x^{(i)}) + f^1_{1_w}((1-\lambda)\bar{P} + \lambda G_\theta(z^{(i)}))]$
  end for
  Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
  Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
  Update θ using SGD by ascending with:
    $\nabla_\theta \frac{1}{m}\sum_{i=1}^m f^1_{1_w}(G_\theta(z^{(i)}))$
  for t=1, ..., n_{D_2} **do**
    Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
    Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
    Update w using SGD by ascending with:
      $\nabla_w \frac{1}{m}\sum_{i=1}^m [f_{2_w}(x^{(i)}) + f^2_{2_w}(G_\theta(z^{(i)}))]$
  end for
  Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
  Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
  Update θ using SGD by ascending with:
    $\nabla_\theta \frac{1}{m}\sum_{i=1}^m f^2_{2_w}(G_\theta(z^{(i)}))$
end while

Algorithm 2 Training algorithm for IGGAN (PDA + PDA).

Require: The number of D_1 iterations n_{D_1}, the number of D_2 iterations n_{D_2}, batchsize $m = 64$, $f_{1_w}(x) = E_{x\sim p_{data}}[D_{1_w}(T_1(x))]$ and $f^1_{1_w} = -E_{x\sim G_\theta}[f^*(D_{1_w}(T_1((x))))]$ determine the objective function of D_{1_w}, $f_{2_w} = E_{x\sim p_{data}}[D_{2_w}(T_2(x))]$ and $f^2_{2_w} = -E_{x\sim G_\theta}[f^*(D_{2_w}(T_2(x)))]$ determine the objective function of D_{2_w}, where T_1 and T_2 are different PDA methods. w and θ are the parameters of the Ds and G, respectively.

while θ has not converged **do**
  for t=1, ..., n_{D_1} **do**
    Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
    Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
    Update w using SGD by ascending with:
      $\nabla_w \frac{1}{m}\sum_{i=1}^m [f_{1_w}(x^{(i)}) + f^1_{1_w}(G_\theta(z^{(i)}))]$
  end for
  Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
  Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
  Update θ using SGD by ascending with:
    $\nabla_\theta \frac{1}{m}\sum_{i=1}^m f^1_{1_w}(G_\theta(z^{(i)}))$
  for t=1, ..., n_{D_2} **do**
    Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
    Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
    Update w using SGD by ascending with:
      $\nabla_w \frac{1}{m}\sum_{i=1}^m [f_{2_w}(x^{(i)}) + f^2_{2_w}(G_\theta(z^{(i)}))]$
  end for
  Samples $\left\{x^{(i)}\right\}_{i=1}^m \sim P_{data}$
  Samples $\left\{z^{(i)}\right\}_{i=1}^m \sim P_z$
  Update θ using SGD by ascending with:
    $\nabla_\theta \frac{1}{m}\sum_{i=1}^m f^2_{2_w}(G_\theta(z^{(i)}))$
end while

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019. 2

[2] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[3] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010. 1, 2

[4] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NeurIPS*, 2016. 1

[5] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *ICLR*, 2021. 1

[6] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. On data augmentation for gan training. *IEEE Transactions on Image Processing*, 30:1882–1897, 2021. 1, 2

[7] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. In *NeurIPS*, 2020. 3