1. Proofs

**Theorem 1.** Theoretical analysis of IGGAN.

**Case 1: IGGAN with NDA in \( D_1 \) and PDA in \( D_2 \)**

Let \( \bar{P} \in \mathcal{P}(\chi) \) be any distribution over \( \chi \) with disjoint support than \( p_{\text{data}} \), such that \( \text{supp}(p_{\text{data}}) \cap \text{supp}(\bar{P}) = \emptyset \). Let \( P^T \in p_{\text{data}} \) be any distribution over real data. Let \( D_1 : \chi \rightarrow \mathbb{R} \) and \( D_2 : \chi \rightarrow \mathbb{R} \) be the set of discriminators over \( \chi \), \( f : \mathbb{R} \geq 0 \rightarrow \mathbb{R} \) be a convex, semi-continuous function such that \( f(1) = 0 \), \( f^* \) be the convex conjugate of \( f \), \( \frac{df}{dx} \) be the derivative of \( f \), \( G_\theta \) be a distribution with sample space \( \chi \), and \( G^T_\theta \in G_\theta \) be any distribution over sample space \( \chi \). \( T \) is one kind of PDA. Then \( \forall \lambda \in (0, 1) \), we have

\[
\begin{align*}
\argmax_{\lambda, \theta} \max_{G_\theta \in \mathcal{P}(\chi)} L_f(G_\theta, D_1, D_2) &= \argmax_{\lambda, \theta} \max_{G_\theta \in \mathcal{P}(\chi)} L_f(\lambda G_\theta + (1 - \lambda) \bar{P}, D_1, D_2) \\
&= \text{argmax}_{G_\theta \in \mathcal{P}(\chi)} \left( \lambda L_f(G_\theta, D_1, D_2) + (1 - \lambda) \bar{P}, D_1, D_2 \right) \\
&= \text{argmax}_{G_\theta \in \mathcal{P}(\chi)} \left( \lambda L_f(G_\theta, D_1, D_2) + (1 - \lambda) \bar{P}, D_1, D_2 \right) \\
&= \text{argmax}_{G_\theta \in \mathcal{P}(\chi)} \left( \lambda L_f(G_\theta, D_1, D_2) + (1 - \lambda) \bar{P}, D_1, D_2 \right)
\end{align*}
\]

(S1)

**Proof.** Let \( p(x), \bar{p}(x), p^T(x) \) and \( q(x) \) denote the density functions of \( p_{\text{data}}, \bar{P}, P^T \) and \( G_\theta \) 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

\[
\begin{align*}
\argmax_{D_1 : \chi \rightarrow \mathbb{R}} L_f(Q, D) &= \hat{f} \left( \frac{p_{\text{data}}}{Q} \right). \\
&= \hat{f} \left( \frac{p_{\text{data}}/(\lambda G_\theta + (1 - \lambda) \bar{P})}{D_1 : \chi \rightarrow \mathbb{R}} \right)
\end{align*}
\]

(S4)

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

\[
\begin{align*}
\argmax_{D_1 : \chi \rightarrow \mathbb{R}} L_f(Q, D) &= \hat{f} \left( \frac{p_{\text{data}}}{\lambda G_\theta + (1 - \lambda) \bar{P}} \right).
\end{align*}
\]

(S5)

\[
\begin{align*}
\argmax_{D_2 : \chi \rightarrow \mathbb{R}} L_f(Q, D) &= \hat{f} \left( \frac{P^T}{G^T_\theta} \right).
\end{align*}
\]

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

\[
\begin{align*}
\argmax_{D_1 : \chi \rightarrow \mathbb{R}} \max_{G_\theta \in \mathcal{P}(\chi)} L_f(G_\theta, D_1, D_2) &= p_{\text{data}}, \\
where \quad L_f(G_\theta, D_1, D_2) &= E_{x \sim p_{\text{data}}} \left[ D_1(T_1(x)) - E_{x \sim G_\theta} \left[ f^* \left( D_1(T_1(x)) \right) \right] - E_{x \sim p_{\text{data}}} \left[ D_2(T_2(x)) \right] - E_{x \sim G_\theta} \left[ f^* \left( D_2(T_2(x)) \right) \right] \right]
\end{align*}
\]

(S7)

**Proof.** Let \( p(x), \bar{p}(x), \bar{p}^T(x) \) and \( q(x) \) denote the density functions of \( p_{\text{data}}, \bar{P}, P^{T_1} \) and \( G_\theta \) respectively (and \( P, \bar{P}, P^{T_1}, P^{T_2} \) and \( G_\theta \) 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 \( D_2 \), following 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.
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.

\[ P_{T1}, P_{T2}, Q \text{ for the respective distributions} \]. 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

\[
\arg\max_{D: X \rightarrow \mathbb{R}} L_f(Q, D) = f'(p_{\text{data}}/Q). \tag{S10}
\]

Therefore, by replacing \( p_{\text{data}} \) with \( P_{T1} \) and \( Q \) with \( G^{T_1}_{\theta} \) in \( D_1 \) and \( p_{\text{data}} \) with \( P_{T2} \) and \( Q \) with \( G^{T_2}_{\theta} \) in \( D_2 \), we have

\[
\arg\max_{D_1: X \rightarrow \mathbb{R}} L_f(G_{\theta}, D_1) = f'(P_{T1}/G^{T_1}_{\theta}). \tag{S11}
\]

\[
\arg\max_{D_2: X \rightarrow \mathbb{R}} L_f(G_{\theta}, D_2) = f'(P_{T2}/G^{T_2}_{\theta}). \tag{S12}
\]

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.

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.
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 $f_{1_w}^* = -E_{x \sim G_\theta}[f(D_{1_w}(x))]$ determine the objective function of $D_{1_w}$, $f_{2_w} = E_{x \sim P_{data}}[D_{2_w}(x)]$ and $f_{2_w}^* = -E_{x \sim G_\theta}[f(D_{2_w}(x))]$ determine the objective function of $D_{2_w}$, where $T$ is one PDA method. $\bar{P}$ is the distribution of the NDA. $w$ and $\theta$ are the parameters of the $D$s and $G$, respectively.

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

References


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 $m = 64$, $f_{1_w}(x) = E_{x \sim P_{data}}[D_{1_w}(x)]$ and $f_{1_w}^* = -E_{x \sim G_\theta}[f(D_{1_w}(x))]$ determine the objective function of $D_{1_w}$, $f_{2_w} = E_{x \sim P_{data}}[D_{2_w}(x)]$ and $f_{2_w}^* = -E_{x \sim G_\theta}[f(D_{2_w}(x))]$ determine the objective function of $D_{2_w}$, where $T_1$ and $T_2$ are different PDA methods. $w$ and $\theta$ are the parameters of the $D$s and $G$, respectively.

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