Supplementary Material

This document presents a detailed discussion on the theorems, proofs, experimental observations and setup left out in the main paper due to space constraints.

1. Background

1.1. Preliminaries and Notations

Consider the GAN game defined by :

$$\min_{\theta_g \in \Theta_G} \max_{\theta_d \in \Theta_D} V(D_{\theta_d}, G_{\theta_g}), \tag{1}$$

where the generator (G, parametrized by θ_g) and discriminator (D, parametrized by θ_d) are neural networks and V is the objective function that the agents seek to optimize. Let us denote by P_r the real data distribution and by P_{θ_g} the generated data distribution. We consider three different GAN formulations, each expressing the objective function (V) as summarized below:

Classic GAN, defined by:

$$V_c = \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_r}[\log D(\mathbf{x})] + \frac{1}{2}\mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[\log(1 - D(\mathbf{x}))] \tag{2}$$

F-GAN, defined by:

$$V_f = \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[f^*(D(\mathbf{x}))], \tag{3}$$

where f^* denotes the Fenchel conjugate of a convex lower semi-continuous function f satisfying $f(1) = 0$.

WGAN, defined by:

$$V_w = \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[D^c(\mathbf{x})], \tag{4}$$

where D^c denotes the $c-$transform of D.

We denote by $DIV(P_{\theta_g}||P_r)$ the divergence between the real and generated data distributions, defined for the three GAN formulations as below:

$$DIV(P_{\theta_g}||P_r) = \begin{cases} JSD(P_{\theta_g}||P_r), & \text{if } V = V_c \\ W_c(P_{\theta_g}||P_r), & \text{if } V = V_w \\ D_f(P_{\theta_g}||P_r), & \text{if } V = V_f \end{cases}$$

where JSD, W_c and D_f denotes the Jenson-Shannon Divergence, Wasserstein Distance (associated with a transport cost c) and $f-$divergence respectively. To theoretically study the properties of DG^λ, we assume that for a fixed generator, an optimal discriminator (D_w) that maximizes V exists.

1.2. GANs Need Not Converge to Nash Equilibrium

In this section, we provide further empirical evidence for the claim that *GANs can produce realistic data even at non-Nash critical points*. Section 3.2 of the main paper presented results for an SNGAN trained over the CIFAR-10 dataset. Figure 1 demonstrates similar results for SNGAN over the MNIST and CELEB-A datasets. We observe that the converged GAN configurations do not exhibit the characteristics of a Nash equilibrium, despite producing high fidelity samples. A Nash equilibrium is optimal for both the agents, thus no agent can deviate from it to unilaterally improve its payoff. As the generator (discriminator) aims to minimize (maximize) the objective function, a Nash equilibrium would constitute a local minima (maxima) for the generator (discriminator). The hessian of the objective function w.r.t the generator (discriminator) would thus be positive (negative) definite. However, as depicted in Figure 1, the hessian of the objective function w.r.t the generator is indefinite as it has both positive as well as negative eigenvalues, indicating that the configuration is not a local minima for the generator and thus not a Nash equilibrium. This is also verified by the visualization (Figure 1 that the generator is able to deviate from the converged configuration on unilaterally optimizing the objective function, attaining a lower loss, but deteriorating the quality of the learned data distribution.

2. Theorems and Proofs

2.1. Classical Duality Gap

Proposition 1. *The duality gap (DG) for a GAN configuration will tend to zero only at a Nash equilibrium and is positive otherwise.*

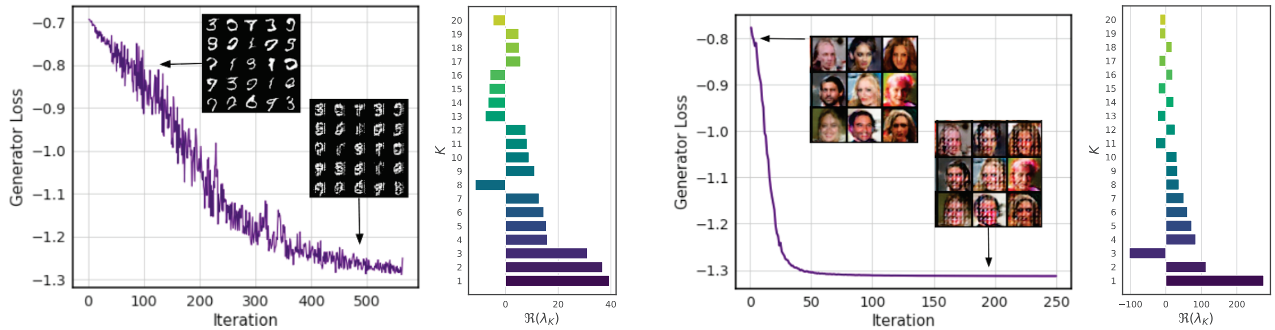
Proof. A configuration (θ_d^*, θ_g^*) of the GAN game (Eq 1) is called a Nash equilibrium if and only if $\forall\, \theta_d, \theta_g$,

$$V(D_{\theta_d}, G_{\theta_g^*}) \leq V(D_{\theta_d^*}, G_{\theta_g^*}) \leq V(D_{\theta_d^*}, G_{\theta_g})$$

Equivalently, $\max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g^*}) = \min_{\tilde{\theta_g} \in \Theta_G} V(D_{\theta_d^*}, G_{\tilde{\theta_g}})$

We have from the definition of duality gap (DG),

$$DG(\theta_d, \theta_g) = \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) - \min_{\tilde{\theta_g} \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta_g}})$$



(a) SNGAN trained on MNIST

(b) SNGAN trained on CELEB-A

Figure 1. The high fidelity images outputted by a converged GAN deteriorates on optimizing only w.r.t the generator while attaining a lower loss, indicating that the GAN has not converged to a Nash equilibrium; confirmed by the presence of positive and negative eigenvalues in the Hessian.

Thus, when $DG(\theta_d, \theta_g) = 0$

$$\implies \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) = \min_{\tilde{\theta_g} \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta_g}})$$

$$\implies (\theta_d, \theta_g) \text{ is a Nash equilibrium.}$$

Similarly, when (θ_d, θ_g) is a Nash equilibrium,

$$\implies \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) = \min_{\tilde{\theta_g} \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta_g}})$$

$$\implies DG(\theta_d, \theta_g) = 0$$

When (θ_d, θ_g) does not constitute a Nash equilibrium,

$$\implies \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) > \min_{\tilde{\theta_g} \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta_g}})$$

$$\implies DG(\theta_d, \theta_g) > 0$$

□

However, as discussed, GANs can converge to non-Nash critical points while producing data samples of high fidelity. The behaviour of DG at such scenarios is not well understood, limiting its applicability as a tool for monitoring GAN training.

2.2. Proximal Duality Gap

Proposition 2. *Consider a GAN game governed by the objective function V. Then, for a configuration (θ_d, θ_g) the proximal objective V^λ is related to V as :*

$$V^\lambda(\theta_d, \theta_g) \leq V_{D_w}(\theta_g)$$

Proof. Since $\lambda ||D_{\tilde{\theta_d}} - D_{\theta_d}||^2 \geq 0$, we have

$$\begin{aligned} V(D_{\tilde{\theta_d}}, G_{\theta_g}) - \lambda ||D_{\theta_d} - D_{\tilde{\theta_d}}||^2 &\leq V(D_{\tilde{\theta_d}}, G_{\theta_g}) \\ \implies V^\lambda(\theta_d, \theta_g) &\leq \max_{\tilde{\theta_d}} V(D_{\tilde{\theta_d}}, G_{\theta_g}) \\ &= V_{D_w}(\theta_g) \end{aligned}$$

□

Lemma 1. *Given a generator θ_g, V_{D_w} is related to the divergences between P_r and P_{θ_g} in the various GAN objectives as follows*

$$V_{D_w}(\theta_g) = \begin{cases} JSD(P_{\theta_g}||P_r) - \log 2, & \text{if } V = V_c \\ W_c(P_{\theta_g}||P_r), & \text{if } V = V_w \\ D_f(P_{\theta_g}||P_r), & \text{if } V = V_f \end{cases}$$

Proof. We divide the proof into three parts corresponding to each of the three GAN formulations.

Part 1. *Classic GAN*
Consider the classic GAN objective $V = V_c$. We have,

$$\begin{aligned} V &= \frac{1}{2}\mathbb{E}_{x\sim P_r}[\log D(x)] + \frac{1}{2}\mathbb{E}_{x\sim P_{\theta_g}}[\log(1 - D(x))] \\ &= \frac{1}{2}\int P_r(x)\log D(x)dx + \frac{1}{2}\int P_{\theta_g}(x)\log(1 - D(x))dx \end{aligned}$$

We have $V_{D_w} = \max_D V$. The worst case discriminator D_w can be obtained by differentiating V w.r.t D for every x and equating to zero. This gives:

$$D_w(x) = \frac{P_r(x)}{P_{\theta_g}(x) + P_r(x)}$$

Substituting D_w back into V gives,

$$\begin{aligned} V_{D_w} =& \frac{1}{2}\int P_r(x)\log\left(\frac{P_r}{P_r + P_{\theta_g}}\right)dx \\ &+ \frac{1}{2}\int P_{\theta_g}(x)\log\left(\frac{P_{\theta_g}}{P_r + P_{\theta_g}}\right)dx \\ =& JSD(P_{\theta_g}||P_r) - \log 2 \end{aligned}$$

Part 2. *Wasserstein GAN*
Consider the Wasserstein GAN objective $V = V_w$. We have,

$$V = \mathbb{E}_{x\sim P_r}[D(x)] - \mathbb{E}_{x\sim P_{\theta_g}}[D^c(x)]$$

where, $D^c(x)$ is related to $D(x)$ as

$$\begin{aligned} D^c(x) &= \sup_{x'} \{ \ D(x') - c(x, x') \ \} \\ &\geq D(x) - c(x, x) \\ &\geq D(x) \end{aligned}$$

Substituting for $D^c(x)$ into V, we have

$$V \leq \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{x \sim P_{\theta_g}}[D(x)]$$

Thus, if $\exists$ a $c-$concave function D_w such that $D_w(x) = D_w^c(x)$, then the bound is attainable and we have,

$$\begin{aligned} V_{D_w} &= \max_{D \ c-concave} V = \sup_{D \ c-concave} V \\ &= \mathbb{E}_{x \sim P_r}[D_w(x)] - \mathbb{E}_{x \sim P_{\theta_g}}[D_w(x)] \end{aligned}$$

Consider a constant discriminator $D_{constant}(x) = k$, which by definition satisfies $c-$concavity. We have,

$$\begin{aligned} D^c_{constant}(x) &= \sup_{x'} \{ \ D_{constant}(x') - c(x, x') \ \} \\ &= \sup_{x'} \{ \ k - c(x, x') \ \} \\ &= k + \sup_{x'} \{ \ -c(x, x') \ \} \\ &= k = D_{constant}(x) \end{aligned}$$

Thus, for $D_w = D_{constant}$ the bound is attainable and we have,

$$\begin{aligned} V_{D_w} &= \max_{D \ c-concave} V \\ &= \sup_{D \ c-concave} \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{x \sim P_{\theta_g}}[D^c(x)] \\ &= W_c(P_{\theta_g} || P_r) \end{aligned}$$

Part 3. *F-GAN*
Consider the F-GAN objective $V = V_f$. We have,

$$V = \mathbb{E}_{x \sim P_r}[D(x)] - \mathbb{E}_{x \sim P_{\theta_g}}[f^*(D(x))],$$

where f^* is the Fenchel conjugate of a convex function f defined by $f^*(x) = \max_t \{ \ xt - f(t) \ \}$. The maximum implied by f^* can be obtained by differentiating it w.r.t t and equating to zero. This gives

$$\begin{aligned} & x - f'(t) = 0 \\ & \implies x = f'(t) \end{aligned}$$

On Substituting the value of x in $f^*(x)$, we get that f^* satisfies the property

$$f^*(f'(t)) = tf'(t) - f(t) \tag{5}$$

We have $V_{D_w} = \max_D V$. The worst case discriminator D_w can be obtained by differentiating V w.r.t D for every x and equating to zero. This gives:

$$D_w(x) = f^{*'-1}\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right)$$

Substituting D_w back into V we get,

$$\begin{aligned} V_{D_w} = &\int P_r(x) f^{*'-1}\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right) dx \\ &- \int P_{\theta_g}(x) f^*\left(f^{*'-1}\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right)\right) dx \end{aligned}$$

The Fenchel conjugate f^* of a convex function f satisfies $f^{*'-1} = f'$. Thus, substituting for $f^{*'-1}$ in V_{D_w}, we get,

$$\begin{aligned} V_{D_w} &= \int P_r(x) f'\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right) dx \\ &\quad - \int P_{\theta_g}(x) f^*\left(f'\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right)\right) dx \\ &= \int P_r(x) f'\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right) dx \\ &\quad - \int P_{\theta_g}(x) \left(\frac{P_r(x)}{P_{\theta_g}(x)} f'\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right)\right) dx \\ &\quad + \int P_{\theta_g}(x) \left(f\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right)\right) dx \quad \text{(using 5)} \\ &= \int P_{\theta_g}(x) f\left(\frac{P_r(x)}{P_{\theta_g}(x)}\right) dx \\ &= D_f(P_{\theta_g} || P_r) \end{aligned}$$

□

Theorem 1. *Consider a GAN game governed by an objective function V. Then the proximal duality gap (DG^λ) at a configuration (θ_d, θ_g) is related to the divergence between the real (P_r) and generated (P_{θ_g}) data distributions as follows.*

$$DG^\lambda(\theta_d, \theta_g) \geq DIV(P_{\theta_g} || P_r) - \kappa$$

where,

$$DIV(P_{\theta_g} || P_r) = \begin{cases} JSD(P_{\theta_g} || P_r), & \textit{if } V = V_c \\ W_c(P_{\theta_g} || P_r), & \textit{if } V = V_w \\ D_f(P_{\theta_g} || P_r), & \textit{if } V = V_f \end{cases}$$

and κ (≥ 0) denotes the minimum divergence that the considered class of generator functions can achieve with the real data distribution.

Proof. We divide the proof into three parts corresponding to each of the the three GAN formulations. For each GAN formulation, we denote by κ the minimum divergence that the considered class of generator functions can attain with the real data distributions; $\min_{P_{\theta_g}} \ DIV(P_{\theta_g}||P_r) = \kappa$. Note that under the realizable setting, $\kappa = 0$ as $\exists\theta_g$ such that $P_{\theta_g} = P_r$.

Part 1. *Classic GAN*,
Consider the classic GAN objective $V = V_c$. We have,
$V_{D_w}(\theta_g) = JSD(P_{\theta_g}||P_r) - \log 2$ (lemma 1)

Further, $V^{\lambda}(\theta_d, \theta_g) \leq V_{D_w}(\theta_g)$ (proposition 2)
$\implies V^{\lambda}(\theta_d, \theta_g) \leq JSD(P_{\theta_g}||P_r) - \log 2$
Thus, using a slight misuse of notation, we have

$$\begin{aligned}
V^{\lambda}_{G_w}(\theta_d) &= \min_{\theta_g'} \ V^{\lambda}(\theta_d, \theta_g^{'}) \\
&\leq \min_{P_{\theta_g'}} \ JSD(P_{\theta_g'}||P_r) - \log 2 \\
&= \kappa - \log 2 \\
\therefore DG^{\lambda}(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V^{\lambda}_{G_w}(\theta_d) \\
&\geq JSD(P_{\theta_g}||P_r) - \kappa
\end{aligned}$$

Part 2. *Wasserstein GAN*,
Consider the Wasserstein GAN objective $V = V_w$. We have, $V_{D_w}(\theta_g) = W_c(P_{\theta_g}||P_r)$ (lemma 1)

Further, $V^{\lambda}(\theta_d, \theta_g) \leq V_{D_w}(\theta_g)$ (proposition 2)
$\implies V^{\lambda}(\theta_d, \theta_g) \leq W_c(P_{\theta_g}||P_r)$
Thus, using a slight misuse of notation, we have

$$\begin{aligned}
V^{\lambda}_{G_w}(\theta_d) &= \min_{\theta_g'} \ V^{\lambda}(\theta_d, \theta_g^{'}) \\
&\leq \min_{P_{\theta_g'}} \ W_c(P_{\theta_g'}||P_r) \\
&= \kappa \\
\therefore DG^{\lambda}(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V^{\lambda}_{G_w}(\theta_d) \\
&\geq W_c(P_{\theta_g}||P_r) - \kappa
\end{aligned}$$

Part 3. *F-GAN*,
Consider the F-GAN objective $V = V_f$. We have, $V_{D_w}(\theta_g) = D_f(P_{\theta_g}||P_r)$ (lemma 1)

Further, $V^{\lambda}(\theta_d, \theta_g) \leq V_{D_w}(\theta_g)$ (proposition 2)
$\implies V^{\lambda}(\theta_d, \theta_g) \leq D_f(P_{\theta_g}||P_r)$
Thus, using a slight misuse of notation, we have

$$\begin{aligned}
V^{\lambda}_{G_w}(\theta_d) &= \min_{\theta_g'} \ V^{\lambda}(\theta_d, \theta_g^{'}) \\
&\leq \min_{P_{\theta_g'}} \ D_f(P_{\theta_g'}||P_r) \\
&= \kappa \\
\therefore DG^{\lambda}(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V^{\lambda}_{G_w}(\theta_d) \\
&\geq D_f(P_{\theta_g}||P_r) - \kappa
\end{aligned}$$

□

Theorem 2. *The proximal duality gap (DG^{λ}) at a configuration (θ_d^*, θ_g^*) for the GAN game defined by V_c, V_w, or V_f is equal to zero for $\lambda = 0$, when the generator learns the real data distribution .i.e, $P_{\theta_g^*} = P_r \implies DG^{\lambda=0}(\theta_d^*, \theta_g^*) = 0$.*

Proof. Let (θ_d^*, θ_g^*) be a GAN configuration such that $P_{\theta_g^*} = P_r$. Since this is a realizable setting, $\min_{P_{\theta_g}} \ DIV(P_{\theta_g}||P_r) = 0$. We divide the proof into three parts corresponding to each of the the three GAN formulations.

Part 1. *Classic GAN*,
Consider the classic GAN objective $V = V_c$. We have,

$$\begin{aligned}
V_{D_w}(\theta_g^*) &= JSD(P_{\theta_g^*}||P_r) - \log 2 \quad \text{(lemma 1)} \\
&= -\log 2 \quad (\because P_{\theta_g^*} = P_r) \\
V^{\lambda=0}_{G_w}(\theta_d^*) &= \min_{\theta_g'} \ V^{\lambda=0}(\theta_d^*, \theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ \max_{\tilde{\theta}_d} \ V(\tilde{\theta}_d, \theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ V_{D_w}(\theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ JSD(P_{\theta_g'}||P_r) - \log 2 \\
&= -\log 2 \\
\therefore DG^{\lambda=0}(\theta_d^*, \theta_g^*) &= V_{D_w}(\theta_g^*) - V^{\lambda=0}_{G_w}(\theta_d^*) \\
&= \ 0
\end{aligned}$$

Part 2. *Wasserstein GAN*,
Consider the Wasserstein GAN objective $V = V_w$. We have,

$$\begin{aligned}
V_{D_w}(\theta_g^*) &= W_c(P_{\theta_g^*}||P_r) \qquad \text{(lemma 1)} \\
&= 0 \quad (\because P_{\theta_g^*} = P_r) \\
V^{\lambda=0}_{G_w}(\theta_d^*) &= \min_{\theta_g'} \ V^{\lambda=0}(\theta_d^*, \theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ \max_{\tilde{\theta_d}} \ V(\tilde{\theta}_d, \theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ V_{D_w}(\theta_g^{'}) \\
&= \min_{P_{\theta_g'}} \ W_c(P_{\theta_g'}||P_r) \\
&= 0 \\
\therefore DG^{\lambda=0}(\theta_d^*, \theta_g^*) &= V_{D_w}(\theta_g^*) - V^{\lambda=0}_{G_w}(\theta_d^*) \\
&= \ 0
\end{aligned}$$

Part 3. *F-GAN*,

Consider the F-GAN objective $V = V_f$. We have,

$$
\begin{aligned}
V_{D_w}(\theta_g^*) &= D_f(P_{\theta_g^*}||P_r) \qquad \text{(lemma 1)}\\
&= 0 \quad (\because P_{\theta_g^*} = P_r)\\
V_{G_w}^{\lambda=0}(\theta_d^*) &= \min_{\theta_g^{'}} \; V^{\lambda=0}(\theta_d^*, \theta_g^{'})\\
&= \min_{P_{\theta_g^{'}}} \; \max_{\tilde{\theta_d}} \; V(\tilde{\theta}_d, \theta_g^{'})\\
&= \min_{P_{\theta_g^{'}}} \; V_{D_w}(\theta_g^{'})\\
&= \min_{P_{\theta_g^{'}}} \; D_f(P_{\theta_g^{'}}||P_r)\\
&= 0\\
\therefore DG^{\lambda=0}(\theta_d^*, \theta_g^*) &= V_{D_w}(\theta_g^*) - V_{G_w}^{\lambda=0}(\theta_d^*)\\
&= \; 0
\end{aligned}
$$

□

Corollary. *For the GAN formulations defined by V_c, V_w or V_f, the generator learns the real data distribution at a configuration (θ_d^*, θ_g^*) if and only if (θ_d^*, θ_g^*) constitutes a Stackelberg equilibrium.*

Proof. Consider a configuration (θ_d^*, θ_g^*) for the GAN game defined by V_c, V_w or V_f. We have,

Part 1. *When $P_{\theta_g^*} = P_r$,*

$$
\begin{aligned}
\text{Theorem 2} &\implies DG^{\lambda=0}(\theta_d^*, \theta_g^*) = 0\\
&\implies (\theta_d^*, \theta_g^*) \in \text{Stackelberg Equilibria}
\end{aligned}
$$

Part 2. *When $(\theta_d^*, \theta_g^*) \in$ Stackelberg Equilibria*,

From the definition of DG^λ and Stackelberg Equilibrium,

$$
\begin{aligned}
&DG^{\lambda=0}(\theta_d^*, \theta_g^*) = 0\\
&\implies DIV(P_{\theta_g^*}||P_r) \le 0 \quad \text{(Theorem 1)}\\
&\implies DIV(P_{\theta_g^*}||P_r) = 0 \quad (\because DIV \ge 0)\\
&\implies P_{\theta_g^*} = P_r
\end{aligned}
$$

□

Theorem 3. *Consider a GAN configuration (θ_d, θ_g). Then, $\forall \lambda^{'} \ge \lambda_0$,*

$$DG^{\lambda=\lambda^{'}}(\theta_d, \theta_g) = 0 \implies DG^{\lambda=\lambda_0}(\theta_d, \theta_g) = 0$$

Proof. We know from the definition of DG^λ that $DG^\lambda(\theta_d, \theta_g) = 0$ is a necessary and sufficient condition for (θ_d, θ_g) to be a $\lambda-$proximal equilibrium i.e. , $DG^\lambda(\theta_d, \theta_g) = 0$ implies that $\forall\, \theta_d^{'}, \theta_g^{'}$,

$$
\begin{aligned}
V(D_{\theta_d^{'}}, G_{\theta_g}) &\le V(D_{\theta_d}, G_{\theta_g})\\
&\le \max_{\tilde{\theta_d} \in \Theta_D} \; V(D_{\tilde{\theta_d}}, G_{\theta_g^{'}}) - \lambda||D_{\tilde{\theta_d}} - D_{\theta_d}||^2
\end{aligned}
$$

and vice-versa.

Now, $\forall \lambda^{'} \ge \lambda_0$, the following holds

$$
\begin{aligned}
&\max_{\tilde{\theta_d} \in \Theta_D} \; V(D_{\tilde{\theta_d}}, G_{\theta_g^{'}}) - \lambda^{'}||D_{\tilde{\theta_d}} - D_{\theta_d}||^2\\
&\qquad \le \max_{\tilde{\theta_d} \in \Theta_D} \; V(D_{\tilde{\theta_d}}, G_{\theta_g^{'}}) - \lambda_0||D_{\tilde{\theta_d}} - D_{\theta_d}||^2
\end{aligned}
$$

Thus, (θ_d, θ_g) is a $\lambda^{'}-$proximal equilibrium $\implies (\theta_d, \theta_g)$ is also a λ_0-proximal equilibrium.

$\therefore DG^{\lambda=\lambda^{'}}(\theta_d, \theta_g) = 0 \implies DG^{\lambda=\lambda_0}(\theta_d, \theta_g) = 0$ □

Theorem 4. *Consider a GAN game governed by an objective function V. For $\lambda > 0$, let V^λ denote the proximal objective defined by $V^\lambda(\theta_d, \theta_g) = max_{\tilde{\theta_d}} V(\tilde{\theta}_d, \theta_g) - \lambda||D_{\tilde{\theta_d}} - D_{\theta_d}||^2$. Then, $\forall\, \epsilon > 0$, $\exists\, \delta > 0$ such that if $||D_{\theta_d} - D_{\tilde{\theta_d}}|| < \delta$, then $DG^\lambda(\theta_d, \theta_g) - DIV(P_{\theta_g}||P_r) < \epsilon$ where,*

$$
DIV(P_{\theta_g}||P_r) = \begin{cases} JSD(P_{\theta_g}||P_r), & \text{if } V = V_c\\ W_c(P_{\theta_g}||P_r), & \text{if } V = V_w\\ D_f(P_{\theta_g}||P_r), & \text{if } V = V_f \end{cases}
$$

Proof. We show that for all $\epsilon > 0$ and $\lambda > 0$, $\delta = \sqrt{\frac{\epsilon}{\lambda}}$ satisfies the claim. We provide the proof for F-GAN. The proof for the other GAN formulations follow on the same lines.

Consider a configuration (θ_d, θ_g) for the GAN game defined by the F-GAN objective $V = V_f$. We have, $V_{D_w}(\theta_g) = D_f(P_{\theta_g}||P_r)$ (lemma 1)

Given that $||D_{\theta_d} - D_{\tilde{\theta_d}}|| < \delta$, we have

$$
\begin{aligned}
DG^{\lambda}(\theta_d, \theta_g) - DIV(P_{\theta_g}||P_r) \\
&= V_{D_w}(\theta_g) - V^{\lambda}_{G_w}(\theta_d) - D_f(P_{\theta_g}||P_r) \\
&= D_f(P_{\theta_g}||P_r) - V^{\lambda}_{G_w}(\theta_d) \\
&\qquad - D_f(P_{\theta_g}||P_r) \\
&= -V^{\lambda}_{G_w}(\theta_d) \\
&= -\min_{\theta_g^{'}} V^{\lambda}(\theta_d, \theta_g^{'}) \\
&= -\min_{\theta_g^{'}} \{\max_{\tilde{\theta_d}} V(\tilde{\theta}_d, \theta_g^{'}) \\
&\qquad\qquad - \lambda||D_{\tilde{\theta_d}} - D_{\theta_d}||^2\} \\
&< -\min_{\theta_g^{'}} \{\max_{\tilde{\theta_d}} V(\tilde{\theta}_d, \theta_g^{'}) - \lambda\delta^2\} \\
&= -\min_{\theta_g^{'}} \{V_{D_w}(\theta_g^{'})\} + \lambda\delta^2 \\
&= -\min_{P_{\theta_g^{'}}} \{D_f(P_{\theta_g^{'}}||P_r)\} + \lambda\delta^2 \\
&= \lambda\delta^2 \\
&= \lambda\left(\sqrt{\frac{\epsilon}{\lambda}}\right)^2 \\
&= \epsilon
\end{aligned}
$$

$\square$

3. DG^{λ} Estimation

Algorithm 1 Proximal Duality Gap $DG^{\lambda}(\theta_d^t, \theta_g^t)$

Input: GAN configuration - (θ_d^t, θ_g^t), data points x_i
function *prox_opt*(θ_d, θ_g) :
  $\tilde{\theta_d} \longleftarrow \theta_d$
  for *j = 0 to T* **do**
    $V^{\lambda} \longleftarrow V(\tilde{\theta}_d, \theta_g) - \frac{\lambda}{n}\sum_{i=1}^{n} ||\nabla_x D_{\tilde{\theta_d}}(x_i) - \nabla_x D_{\theta_d}(x_i)||_2^2$
    $\tilde{\theta_d} \longleftarrow \tilde{\theta_d} + \eta\nabla_{\tilde{\theta_d}} V^{\lambda}$
  end
  return $\tilde{\theta_d},\ V^{\lambda}$

$\theta_d^w \longleftarrow \theta_d^t$; $\theta_g^w \longleftarrow \theta_g^t$
for *i = 0 to N_ITER* **do**
  $\theta_d^w \longleftarrow \theta_d^w + \eta\nabla_{\theta_d^w} V(\theta_d^w, \theta_g^t)$
  $\theta_d^*, V^{\lambda} \longleftarrow$*prox_opt*$(\theta_d^t, \theta_g^w)$
  $\theta_g^w \longleftarrow \theta_g^w - \eta\nabla_{\theta_g^w} V(\theta_d^*, \theta_g^w)$
end
$V_{D_w} \longleftarrow V(\theta_g^t, \theta_d^w)$
$\theta_d^*, V^{\lambda}_{G_w} \longleftarrow$ *prox_opt*(θ_d^t, θ_g^w)
return $DG^{\lambda}(\theta_d^t, \theta_g^t) = V_{D_w} - V^{\lambda}_{G_w}$

Algorithm 1 summarizes the estimation process for proximal duality gap. Given a configuration (θ_d, θ_g) of the GAN, we estimate V_{D_w} and $V^{\lambda}_{G_w}$ by optimizing the objective function w.r.t the individual agents using gradient descent. We have the proximal objective defined by,

$$V^{\lambda}(D_{\theta_d}, G_{\theta_g}) = \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) - \lambda||D_{\tilde{\theta_d}} - D_{\theta_d}||^2$$

Following (Farnia Ozdaglar, 2020) we use the Sobolev norm in V^{λ}, given by

$$||D|| = \sqrt{\mathbb{E}_{\mathbf{x}\sim P_r}\left[||\nabla_{\mathbf{x}} D(\mathbf{x})||_2^2\right]}$$

The estimation process for DG^{λ} is similar to that of DG, except that the worst case generator for a given discriminator is computed w.r.t the proximal objective (V^{λ}). As depicted in Algorithm 1, the function *prox_opt* uses gradient ascent to estimate the proximal objective V^{λ}. Since λ restricts the neighbourhood within which the discriminator is optimal in V^{λ}, the search space for the optimal discriminator increases as λ decreases. Correspondingly, estimating V^{λ} demands a larger number of gradient steps and becomes computationally infeasible as $\lambda \to 0$.

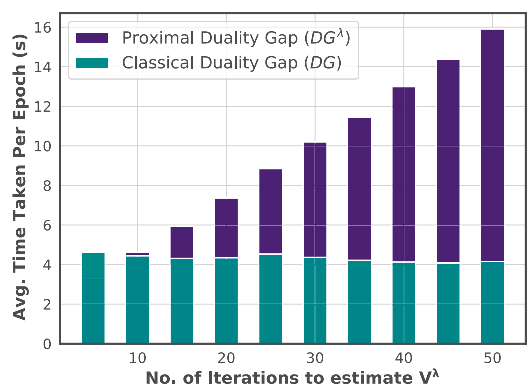


Figure 2. Computational Complexity of DG^{λ} over DG

We thus experimentally studied the computational overhead in estimating DG^{λ} over DG. Figure 2 compares the average time taken per epoch to estimate DG and DG^{λ} across varying gradient steps (T) to approximate V^{λ}. We observe that while DG^{λ} has comparable computational complexity as DG for smaller values of T, it increases rapidly for larger values of T. We observed that for $\lambda = 0.1$, ≈ 20 steps were sufficient for V^{λ} to converge (in line with the observations of (Farnia Ozdaglar, 2020)), incurring computational expense (Figure 2) comparable to that demanded by DG.

Pearson Correlation Coefficient (r)				
	$\mathrm{r}_{DG,IS}$	$\mathrm{r}_{DG^{\lambda},IS}$	$\mathrm{r}_{DG,FID}$	$\mathrm{r}_{DG^{\lambda},FID}$
MNIST	-0.104	**-0.516**	0.207	**0.942**
CIFAR-10	-0.368	**-0.463**	0.293	**0.661**
CELEB-A	-0.535	**-0.846**	0.638	**0.929**

Table 1. Comparing the correlation of DG and DG^{λ} with IS and FID computed during the training of SNGAN over the 3 datasets.

4. Experiments and Results

4.1. Monitoring GAN Training Using DG^{λ}

In this section, we present further empirical observations and evidence to demonstrate the proficiency of proximal duality gap. Section 5.1 of the main paper presented the adeptness of DG^{λ} over DG to monitor GAN convergence, suggesting that the GANs have attained a proximal equilibrium and not a Nash equilibrium. We thus studied the behaviour of the converged GAN while estimating DG and DG^{λ}. We visualized the variation in the generated data distribution across epochs during the estimation of V_{G_w} (for DG) and $V^{\lambda}_{G_w}$ (for DG^{λ}). The behaviours are demonstrated in Figure 3 for SNGAN and Figure 4 for WGAN across the three datasets - (a) MNIST, (b) CIFAR-10 and (c) CELEB-A. The high fidelity of generated data samples depicted in each subfigure suggest that the GANs have converged. However, on optimizing the objective function (V) unilaterally w.r.t the generator, it deviates (top row on the right) and deteriorates the quality of the learned data distribution, validating that the GAN has not attained a Nash equilibrium. However, the generator is unable to deviate (bottom row on the right) from the converged configuration w.r.t the proximal objective (V^{λ}) as the generated data distribution does not vary, indicating that the GANs have attained a proximal equilibrium. This explains why DG^{λ} is able to better characterize convergence over DG for each GAN.

Table 1 presents the correlation of DG and DG^{λ} with the popular quality evaluation measures IS and FID, across the training process of an SNGAN over each of the three datasets. We observe that DG^{λ} has a higher correlation over DG with each of the measures. Thus, further validating the claim that DG^{λ} is adept to not only monitor convergence of the GAN game to an equilibrium, but also the goodness of the learned data distribution.

4.2. Implementation and Hyperparameter Details

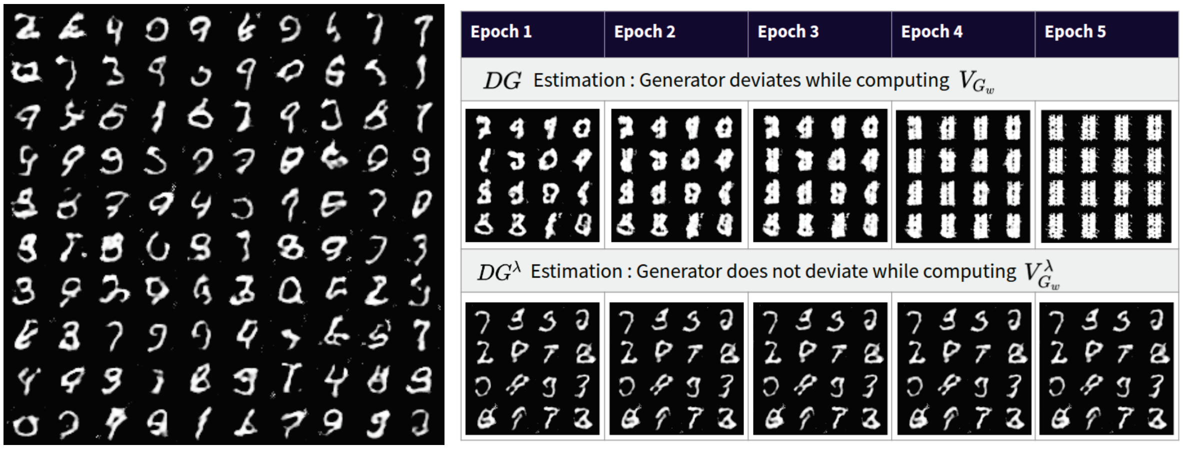
We used the 4-layer DCGAN architecture for both the generator and the discriminator networks in all the experiments. We used an Adam optimizer to train and evaluate all the models. To compute DG^{λ}, we used $\lambda = 0.1$ and 20 optimization steps for approximating the proximal objective. To enforce the Lipschitz constraint in WGAN, we used weight clipping in the range $[-0.01, 0.01]$. We used a batch size of $512, 512, 128$ for MNIST, CIFAR-10 and CELEB-A datasets respectively, where the input images were resized to be of shape 32×32. The latent space dimension for the generator was set to 100. To ensure that we obtain an unbiased estimate for DG^{λ} and DG, we split each dataset into 3 disjoint sets - S_A, S_B and S_C. We trained the GAN using S_A, we used S_B to find the worst case counter parts D_w and G_w via gradient descent, and S_C to evaluate the objective function at the obtained worst case configurations. For each dataset, we kept 5000 samples each in S_B, S_C and the rest in S_A. To estimate the worst case configurations, we optimized each agent unilarerally for 10 epochs over S_B. The learning rates for the discriminator (LR_D) and generator (LR_G), the value in multiples of which the DCGAN architecture steps up the convolutional features ($Step\ Channels$) and the values of (β_1, β_2) used in the Adam optimizer for each of the datasets are summarized in Table 3 for WGAN and Table 2 for SNGAN.

SNGAN					
	LR_D	LR_G	$Step\ Channels$	β_1	β_2
MNIST	$1e-4$	$2e-4$	16	0.00	0.999
CIFAR-10	$1e-4$	$2e-4$	64	0.00	0.999
CELEB-A	$1e-4$	$2e-4$	64	0.00	0.999

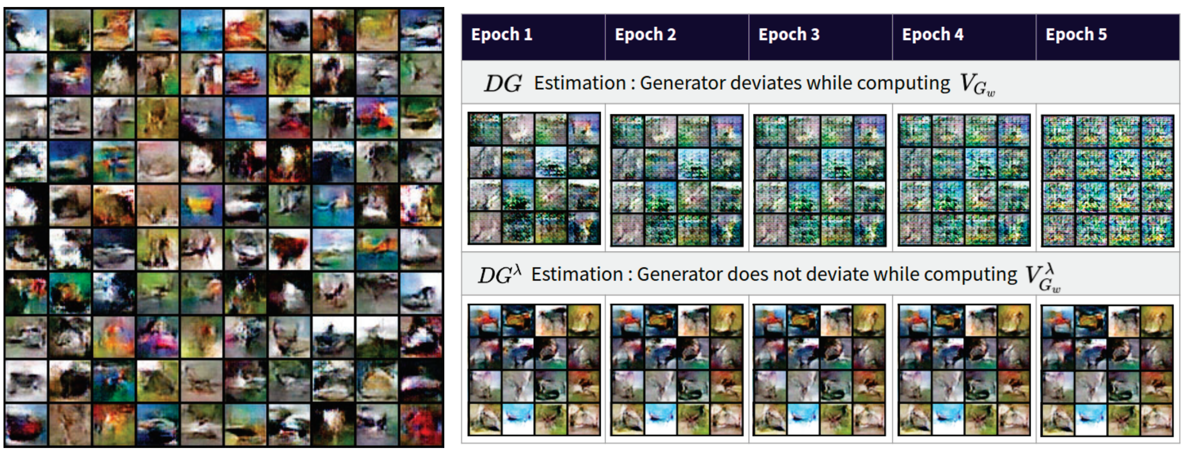
Table 2. Hyperparameter values for SNGAN experiments.

WGAN					
	LR_D	LR_G	$Step\ Channels$	β_1	β_2
MNIST	$4e-4$	$1e-4$	16	0.50	0.999
CIFAR-10	$4e-4$	$1e-4$	64	0.50	0.999
CELEB-A	$4e-4$	$1e-4$	64	0.50	0.999

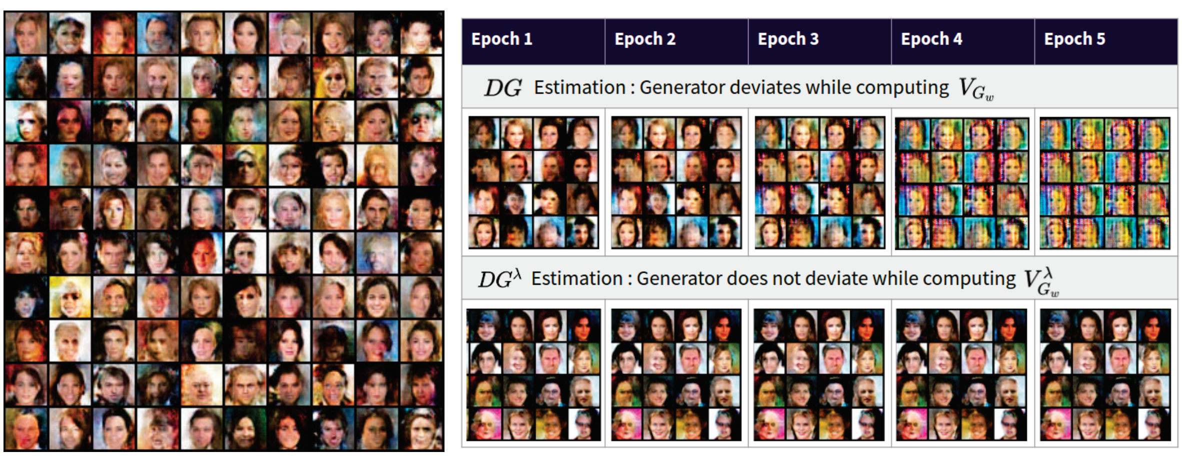
Table 3. Hyperparameter values for WGAN experiments.



(a) MNIST

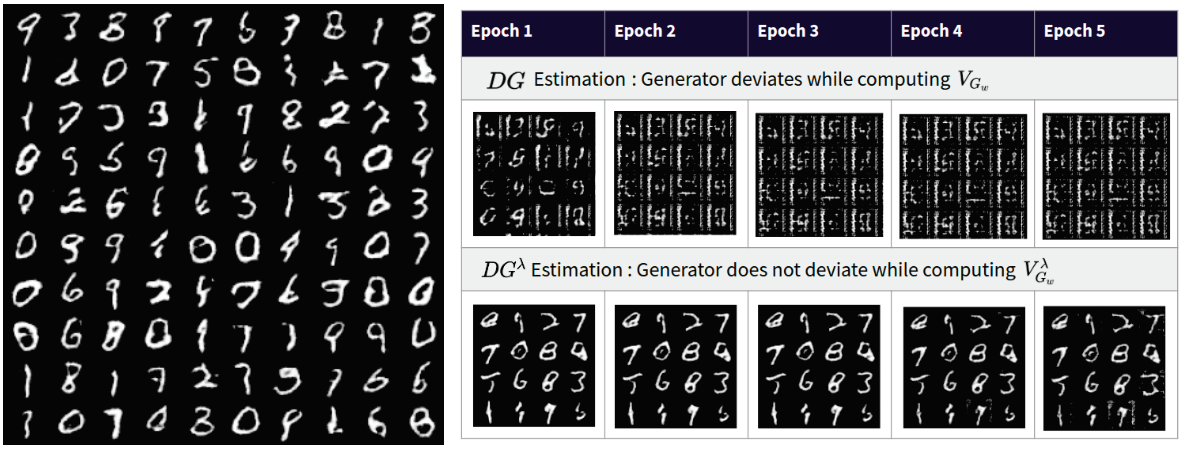


(b) CIFAR-10

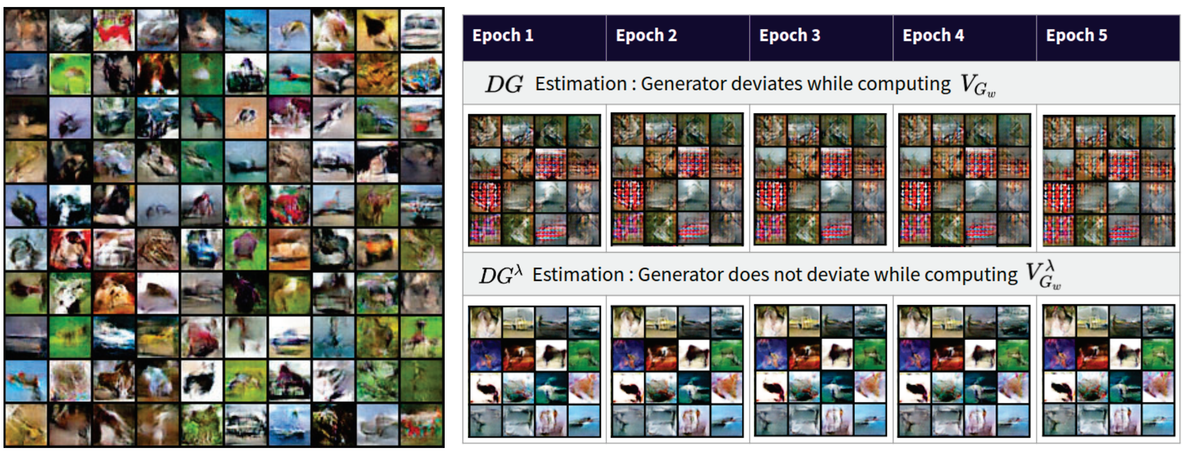


(c) CELEB-A

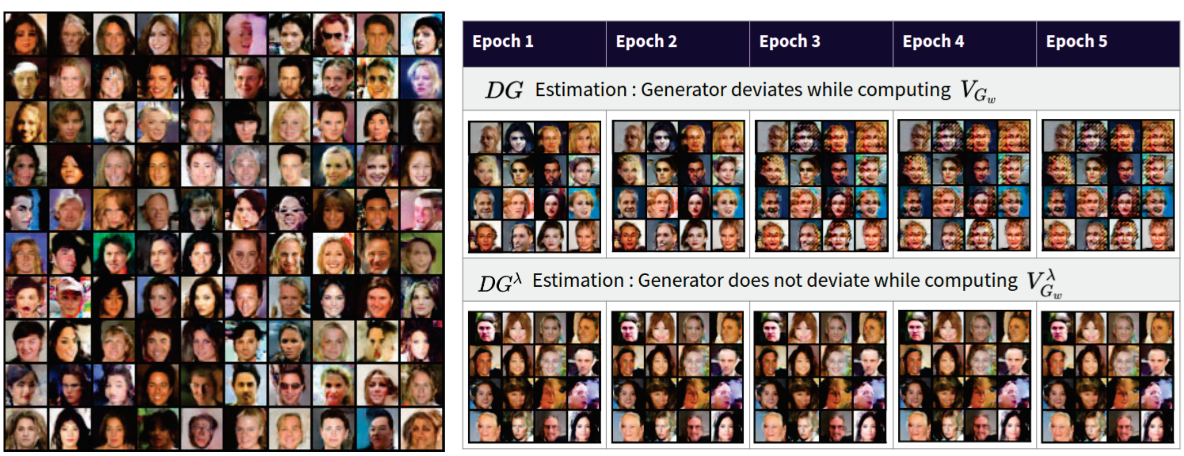
Figure 3. Visualizing the behaviour the generator while estimating DG and DG^{λ} for a converged Wasserstein GAN (WGAN) over the datasets- (a) MNIST (b) CIFAR-10 (c) CELEB-A. The GAN has converged, validated by the high fidelity of generated data samples(left grid). However, the generator deviates on optimizing V, but remains stationary on optimizing V^{λ}, indicating that the converged configuration is a proximal equilibrium and not a Nash equilibrium.



(a) MNIST



(b) CIFAR-10



(c) CELEB-A

Figure 4. Visualizing the behaviour the generator while estimating DG and DG^{λ} for a converged Spectral Normalized GAN (SNGAN) over the datasets- (a) MNIST (b) CIFAR-10 (c) CELEB-A. The GAN has converged, validated by the high fidelity of generated data samples(left grid). However, the generator deviates on optimizing V, but remains stationary on optimizing V^{λ}, indicating that the converged configuration is a proximal equilibrium and not a Nash equilibrium.