

---

## 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 :

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

where the generator ( $G$ , parametrized by  $\theta_g$ ) and discriminator ( $D$ , parametrized by  $\theta_d$ ) 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  $P_{\theta_g}$  the generated data distribution. We consider three different GAN formulations, each expressing the objective function ( $V$ ) as summarized below:

*Classic GAN*, defined by:

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

*F-GAN*, defined by:

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

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

*WGAN*, defined by:

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

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  $DG^\lambda$ , we assume that the real data distribution is learnable. i.e.  $\min_{P_{\theta_g}} DIV(P_{\theta_g} || P_r) = 0$ .

#### 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  $(\theta_d^*, \theta_g^*)$  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),

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

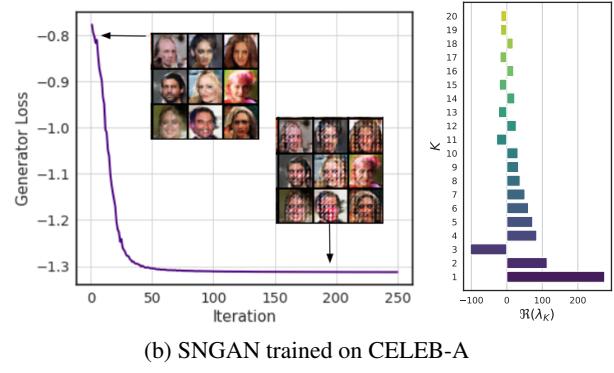
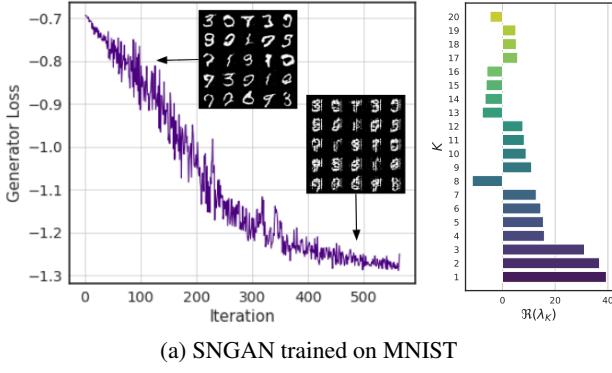


Figure 1. The high fidelity images outputted by a converged GAN deteriorates on optimizing only w.r.t the generator while attaining a lower loss, indicating that the GAN has not converged to a Nash equilibrium; confirmed by the presence of positive and negative eigenvalues in the Hessian.

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

$$\begin{aligned} &\implies \max_{\theta_d \in \Theta_D} V(D_{\tilde{\theta}_d}, G_{\theta_g}) = \min_{\theta_g \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta}_g}) \\ &\implies (\theta_d, \theta_g) \text{ is a Nash equilibrium.} \end{aligned}$$

Similarly, when  $(\theta_d, \theta_g)$  is a Nash equilibrium,

$$\begin{aligned} &\implies \max_{\theta_d \in \Theta_D} V(D_{\tilde{\theta}_d}, G_{\theta_g}) = \min_{\theta_g \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta}_g}) \\ &\implies DG(\theta_d, \theta_g) = 0 \end{aligned}$$

When  $(\theta_d, \theta_g)$  does not constitute a Nash equilibrium,

$$\begin{aligned} &\implies \max_{\theta_d \in \Theta_D} V(D_{\tilde{\theta}_d}, G_{\theta_g}) > \min_{\theta_g \in \Theta_G} V(D_{\theta_d}, G_{\tilde{\theta}_g}) \\ &\implies DG(\theta_d, \theta_g) > 0 \end{aligned}$$

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

$$V_{D_w}(\theta_g) = \begin{cases} 2JSD(P_{\theta_g} || P_r) - \log 4, & \text{if } V = V_c \\ W_c(P_{\theta_g} || P_r), & \text{if } V = V_w \\ D_f(P_{\theta_g} || P_r), & \text{if } V = V_f \end{cases}$$

*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,

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

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,

$$\begin{aligned} V_{D_w} &= \int P_r(x) \log \left( \frac{P_r}{P_r + P_{\theta_g}} \right) dx \\ &\quad + \int P_{\theta_g}(x) \log \left( \frac{P_{\theta_g}}{P_r + P_{\theta_g}} \right) dx \\ &= 2JSD(P_{\theta_g} || P_r) - \log 4 \end{aligned}$$

#### Part 2. Wasserstein GAN

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

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

110 where,  $D^c(x)$  is related to  $D(x)$  as  
 111  
 112  
 113

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

117 Substituting for  $D^c(x)$  into  $V$ , we have  
 118  
 119

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

121 Thus, if  $\exists$  a  $c$ -concave function  $D_w$  such that  $D_w(x) =$   
 122  $D_w^c(x)$ , then the bound is attainable and we have,  
 123

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

124 Consider a constant discriminator  $D_{constant}(x) = k$ , which  
 125 by definition satisfies  $c$ -concavity. We have,  
 126

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

138 Thus, for  $D_w = D_{constant}$  the bound is attainable and we  
 139 have,  
 140

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

### Part 3. F-GAN

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

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

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

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

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

$$f^*(f'(t)) = tf'(t) - f(t) \quad (5)$$

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,

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

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.** The proximal duality gap ( $DG^\lambda$ ) at a configuration  $(\theta_d, \theta_g)$  for the GAN game defined by  $V_c$ ,  $V_w$ , or  $V_f$  is lower bounded by the JSD, Wasserstein distance, and  $f$ -divergence between the real ( $P_r$ ) and generated ( $P_{\theta_g}$ ) data distributions respectively. i.e.,

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

*Proof.* We 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,  
 $V_{D_w}(\theta_g) = 2JSD(P_{\theta_g} || P_r) - \log 4$  (lemma 1)

Further,  $V^\lambda(\theta_d, \theta_g) \leq V_{D_w}(\theta_g)$  (proposition 2)  
 $\implies V^\lambda(\theta_d, \theta_g) \leq 2JSD(P_{\theta_g} || P_r) - \log 4$

165 Thus, using a slight misuse of notation, we have

$$\begin{aligned} V_{G_w}^\lambda(\theta_d) &= \min_{\theta'_g} V^\lambda(\theta_d, \theta'_g) \\ &\leq \min_{P_{\theta'_g}} 2JSD(P_{\theta'_g} || P_r) - \log 4 \\ &= -\log 4 \\ \therefore DG^\lambda(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V_{G_w}^\lambda(\theta_d) \\ &\geq JSD(P_{\theta_g} || P_r) \end{aligned}$$

### 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

$$\begin{aligned} V_{G_w}^\lambda(\theta_d) &= \min_{\theta'_g} V^\lambda(\theta_d, \theta'_g) \\ &\leq \min_{P_{\theta'_g}} W_c(P_{\theta'_g} || P_r) \\ &= 0 \\ \therefore DG^\lambda(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V_{G_w}^\lambda(\theta_d) \\ &\geq W_c(P_{\theta_g} || P_r) \end{aligned}$$

### 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

$$\begin{aligned} V_{G_w}^\lambda(\theta_d) &= \min_{\theta'_g} V^\lambda(\theta_d, \theta'_g) \\ &\leq \min_{P_{\theta'_g}} D_f(P_{\theta'_g} || P_r) \\ &= 0 \\ \therefore DG^\lambda(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V_{G_w}^\lambda(\theta_d) \\ &\geq D_f(P_{\theta_g} || P_r) \end{aligned}$$

□

**Theorem 2.** *The proximal duality gap ( $DG^\lambda$ ) at a configuration  $(\theta_d^*, \theta_g^*)$  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  $(\theta_d^*, \theta_g^*)$  be a GAN configuration such that  $P_{\theta_g^*} = P_r$ . 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,

$$\begin{aligned} V_{D_w}(\theta_g^*) &= 2JSD(P_{\theta_g^*} || P_r) - \log 4 \quad (\text{lemma 1}) \\ &= -\log 4 \quad (\because P_{\theta_g^*} = P_r) \\ V_{G_w}^{\lambda=0}(\theta_d^*) &= \min_{\theta'_g} V^{\lambda=0}(\theta_d^*, \theta'_g) \\ &= \min_{P_{\theta'_g}} \max_{\theta_d} V(\tilde{\theta}_d, \theta'_g) \\ &= \min_{P_{\theta'_g}} V_{D_w}(\theta'_g) \\ &= \min_{P_{\theta'_g}} 2JSD(P_{\theta'_g} || P_r) - \log 4 \\ &= -\log 4 \\ \therefore DG^{\lambda=0}(\theta_d^*, \theta_g^*) &= V_{D_w}(\theta_g^*) - V_{G_w}^{\lambda=0}(\theta_d^*) \\ &= 0 \end{aligned}$$

### Part 2. Wasserstein GAN,

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

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

### Part 3. F-GAN,

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

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

239  $\square$

240 **Corollary.** For the GAN formulations defined by  $V_c$ ,  $V_w$   
 241 or  $V_f$ , the generator learns the real data distribution at a  
 242 configuration  $(\theta_d^*, \theta_g^*)$  if and only if  $(\theta_d^*, \theta_g^*)$  constitutes a  
 243 Stackelberg equilibrium.  
 244

245 *Proof.* Consider a configuration  $(\theta_d^*, \theta_g^*)$  for the GAN game  
 246 defined by  $V_c$ ,  $V_w$  or  $V_f$ . We have,

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

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

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

249 From the definition of  $DG^\lambda$  and Stackelberg Equilibrium,

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

250  $\square$

251 **Theorem 3.** Consider a GAN configuration  $(\theta_d, \theta_g)$ . Then,

252  $\forall \lambda' \geq \lambda_0$ ,

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

253 *Proof.* We know from the definition of  $DG^\lambda$  that  
 254  $DG^\lambda(\theta_d, \theta_g) = 0$  is a necessary and sufficient condition  
 255 for  $(\theta_d, \theta_g)$  to be a  $\lambda$ -proximal equilibrium i.e.,

256  $DG^\lambda(\theta_d, \theta_g) = 0$  implies that  $\forall \theta'_d, \theta'_g$ ,

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

257 and vice-versa.  
 258 Now,  $\forall \lambda' \geq \lambda_0$ , the following holds

$$\begin{aligned}
 \max_{\tilde{\theta}_d \in \Theta_D} V(D_{\tilde{\theta}_d}, G_{\theta'_g}) - \lambda' \|D_{\tilde{\theta}_d} - D_{\theta_d}\|^2 \\
 \leq \max_{\tilde{\theta}_d \in \Theta_D} V(D_{\tilde{\theta}_d}, G_{\theta'_g}) - \lambda_0 \|D_{\tilde{\theta}_d} - D_{\theta_d}\|^2
 \end{aligned}$$

259 Thus,  $(\theta_d, \theta_g)$  is a  $\lambda'$ -proximal equilibrium  $\implies (\theta_d, \theta_g)$   
 260 is also a  $\lambda_0$ -proximal equilibrium.

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

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

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

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

271 Consider a configuration  $(\theta_d, \theta_g)$  for the GAN game de-  
 272 fined by the F-GAN objective  $V = V_f$ . We have,  
 273  $V_{D_w}(\theta_g) = D_f(P_{\theta_g} || P_r)$  (lemma 1)

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

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

□

### 3. $DG^\lambda$ Estimation

---

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


---

Input: GAN configuration -  $(\theta_d^t, \theta_g^t)$ , data points  $x_i$

**function**  $prox\_opt(\theta_d, \theta_g)$ :

```

 $\tilde{\theta}_d \leftarrow \theta_d$ 
for  $j = 0$  to  $T$  do
     $V^\lambda \leftarrow V(\tilde{\theta}_d, \theta_g) - \frac{\lambda}{n} \sum_{i=1}^n \|\nabla_x D_{\tilde{\theta}_d}(x_i) - \nabla_x D_{\theta_d}(x_i)\|_2^2$ 
     $\tilde{\theta}_d \leftarrow \theta_d + \eta \nabla_{\theta_d} V^\lambda$ 
end
return  $\tilde{\theta}_d, V^\lambda$ 

```

$\theta_d^w \leftarrow \theta_d^t; \theta_g^w \leftarrow \theta_g^t$

**for**  $i = 0$  to  $N\_ITER$  **do**

```

 $\theta_d^w \leftarrow \theta_d^w + \eta \nabla_{\theta_d^w} V(\theta_d^w, \theta_g^t)$ 
 $\theta_d^*, V^\lambda \leftarrow prox\_opt(\theta_d^t, \theta_g^w)$ 
 $\theta_g^w \leftarrow \theta_g^w - \eta \nabla_{\theta_g^w} V(\theta_d^*, \theta_g^w)$ 

```

**end**

$V_{D_w} \leftarrow V(\theta_g^t, \theta_d^w)$

$\theta_d^*, V_{G_w}^\lambda \leftarrow prox\_opt(\theta_d^t, \theta_g^w)$

**return**  $DG^\lambda(\theta_d^t, \theta_g^t) = V_{D_w} - V_{G_w}^\lambda$

---

Algorithm 1 summarizes the estimation process for proximal duality gap. Given a configuration  $(\theta_d, \theta_g)$  of the GAN, we estimate  $V_{D_w}$  and  $V_{G_w}^\lambda$  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  $V^\lambda$ , given by

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

The estimation process for  $DG^\lambda$  is similar to that of  $DG$ , except that the worst case generator for a given discriminator is computed w.r.t the proximal objective ( $V^\lambda$ ). As depicted in Algorithm 1, the function  $prox\_opt$  uses gradient ascent to estimate the proximal objective  $V^\lambda$ . Since  $\lambda$  restricts the neighbourhood within which the discriminator is optimal in  $V^\lambda$ , the search space for the optimal discriminator increases as  $\lambda$  decreases. Correspondingly, estimating  $V^\lambda$  demands a larger no. of gradient steps and becomes computationally infeasible as  $\lambda \rightarrow 0$ .

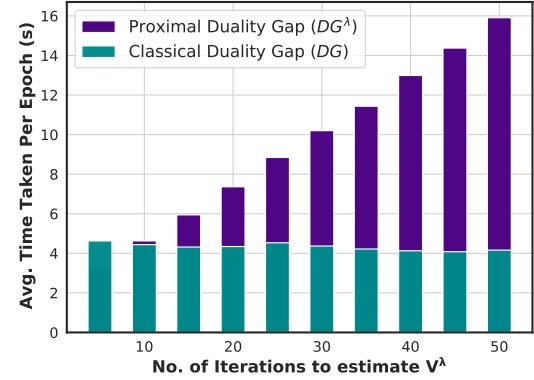


Figure 2. Computational Complexity of  $DG^\lambda$  over  $DG$

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

<b>Pearson Correlation Coefficient (r)</b>				
	$r_{DG,IS}$	$r_{DG^\lambda,IS}$	$r_{DG,FID}$	$r_{DG^\lambda,FID}$
<i>MNIST</i>	-0.104	<b>-0.516</b>	0.207	<b>0.942</b>
<i>CIFAR-10</i>	-0.368	<b>-0.463</b>	0.293	<b>0.661</b>
<i>CELEB-A</i>	-0.535	<b>-0.846</b>	0.638	<b>0.929</b>

Table 1. Comparing the correlation of  $DG$  and  $DG^\lambda$  with IS and FID computed during the training of SNGAN over the 3 datasets.

## 4. Experiments and Results

### 4.1. Monitoring GAN Training Using $DG^\lambda$

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  $DG^\lambda$  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  $DG^\lambda$ . We visualized the variation in the generated data distribution across epochs during the estimation of  $V_{G_w}$  (for  $DG$ ) and  $V_{G_w}^\lambda$  (for  $DG^\lambda$ ). 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 ( $V^\lambda$ ) as the generated data distribution does not vary, indicating that the GANs have attained a proximal equilibrium. This explains why  $DG^\lambda$  is able to better characterize convergence over  $DG$  for each GAN.

Table 1 presents the correlation of  $DG$  and  $DG^\lambda$  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  $DG^\lambda$  has a higher correlation over  $DG$  with each of the measures. Thus, further validating the claim that  $DG^\lambda$  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  $DG^\lambda$ , we used  $\lambda = 0.1$  and

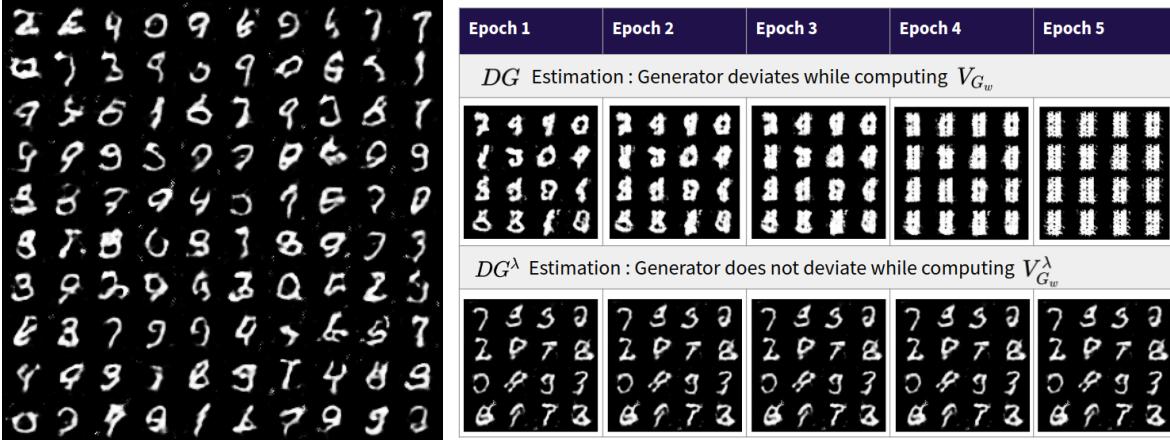
20 optimization steps for approximating the proximal objective. To enforce the Lipschitz constraint in WGAN, we used weight clipping in the range  $[-0.01, 0.01]$ . We used a batch size of 512, 512, 128 for MNIST, CIFAR-10 and CELEB-A datasets respectively, where the input images were resized to be of shape  $32 \times 32$ . The latent space dimension for the generator was set to 100. To ensure that we obtain an unbiased estimate for  $DG^\lambda$  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 unilaterally 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  $(\beta_1, \beta_2)$  used in the Adam optimizer for each of the datasets are summarized in Table 3 for WGAN and Table 2 for SNGAN.

<b>SNGAN</b>					
	$LR_D$	$LR_G$	<i>Step Channels</i>	$\beta_1$	$\beta_2$
<i>MNIST</i>	$1e-4$	$2e-4$	16	0.00	0.999
<i>CIFAR-10</i>	$1e-4$	$2e-4$	64	0.00	0.999
<i>CELEB-A</i>	$1e-4$	$2e-4$	64	0.00	0.999

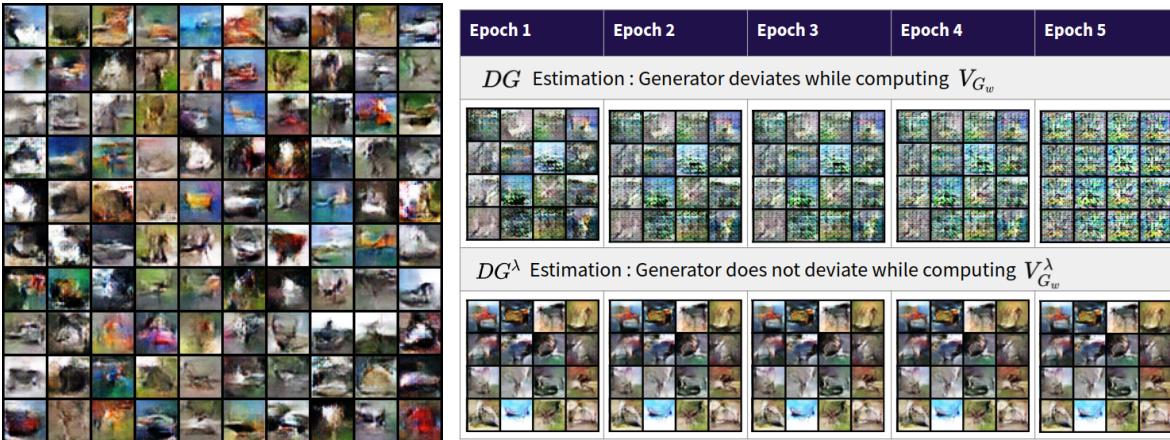
Table 2. Hyperparameter values for SNGAN experiments.

<b>WGAN</b>					
	$LR_D$	$LR_G$	<i>Step Channels</i>	$\beta_1$	$\beta_2$
<i>MNIST</i>	$4e-4$	$1e-4$	16	0.50	0.999
<i>CIFAR-10</i>	$4e-4$	$1e-4$	64	0.50	0.999
<i>CELEB-A</i>	$4e-4$	$1e-4$	64	0.50	0.999

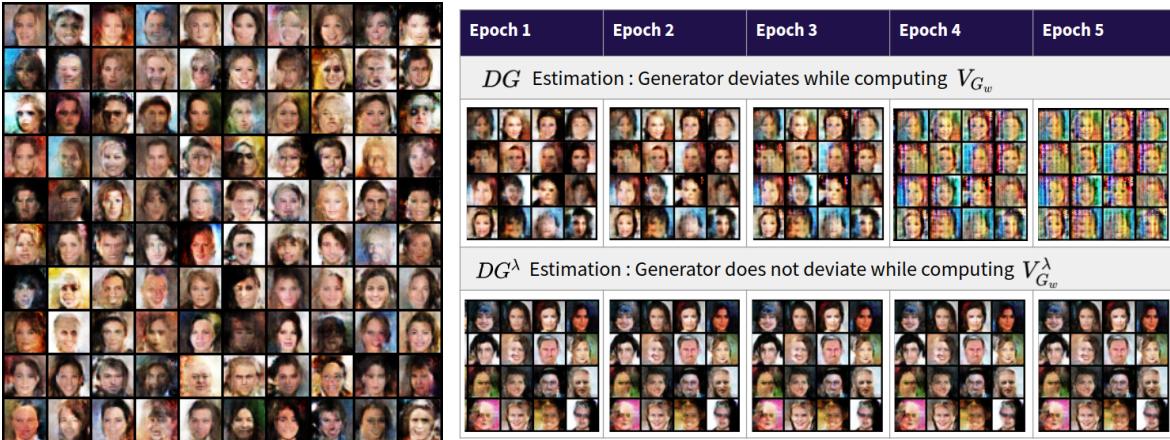
Table 3. Hyperparameter values for WGAN experiments.



(a) MNIST

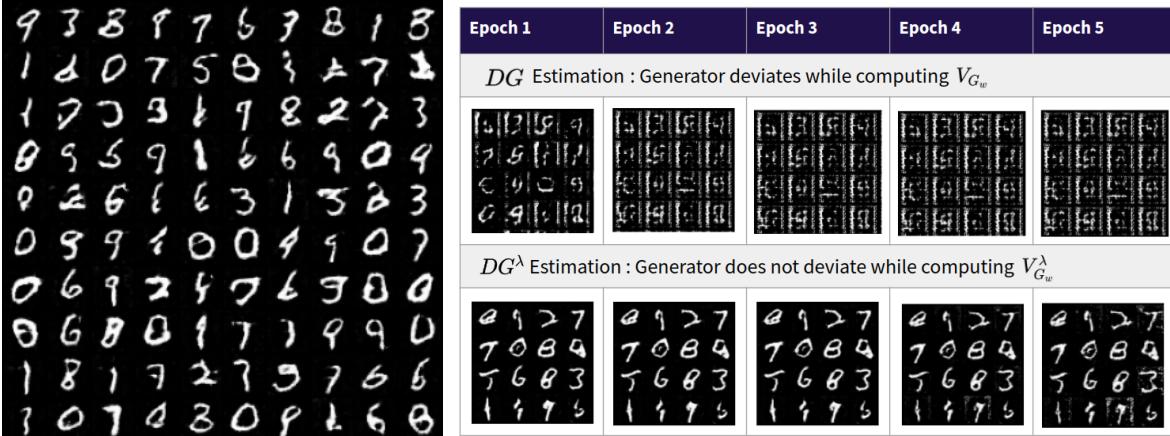


(b) CIFAR-10

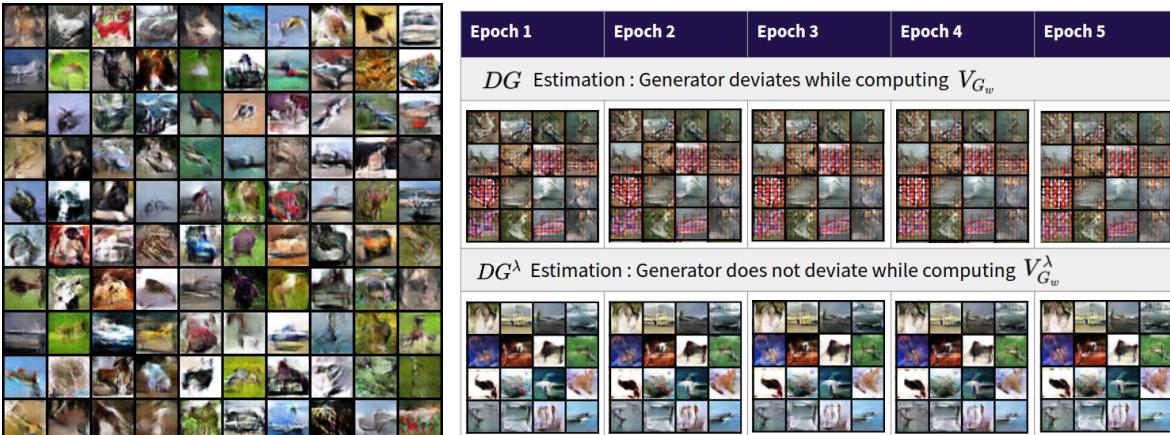


(c) CELEB-A

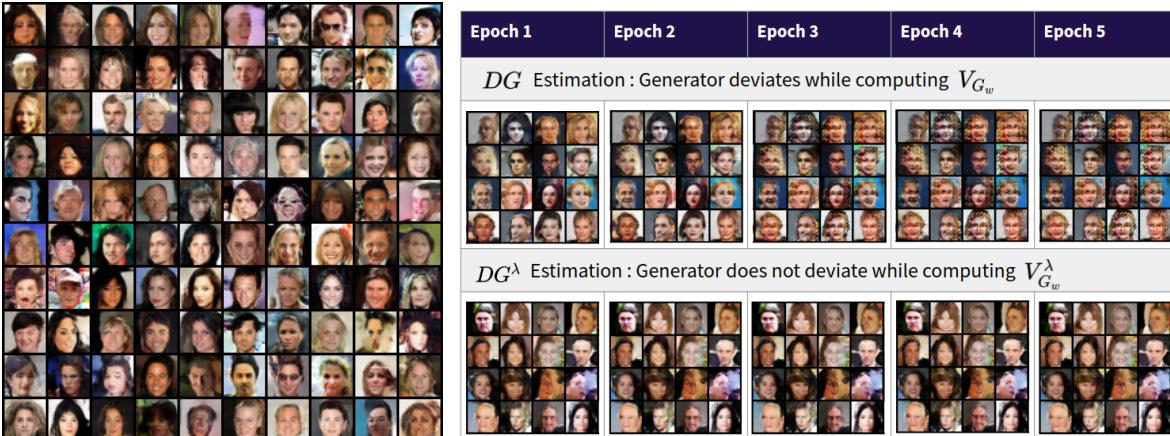
*Figure 3.* Visualizing the behaviour the generator while estimating  $DG$  and  $DG^\lambda$  for a converged Wasserstein GAN (WGAN) over the datasets- (a) MNIST (b) CIFAR-10 (c) CELEB-A. The GAN has converged, validated by the high fidelity of generated data samples(left grid). However, the generator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