Characterizing GAN Convergence Through Proximal Duality Gap

Sahil Sidheekh [1] **Aroof Aimen** [1] **Narayanan C. Krishnan** [1]

Abstract

Despite the accomplishments of Generative Adversarial Networks (GANs) in modeling data distributions, training them remains a challenging task. A contributing factor to this difficulty is the non-intuitive nature of the GAN loss curves, which necessitates a subjective evaluation of the generated output to infer training progress. Recently, motivated by game theory, duality gap has been proposed as a domain agnostic measure to monitor GAN training. However, it is restricted to the setting when the GAN converges to a Nash equilibrium. But GANs need not always converge to a Nash equilibrium to model the data distribution. In this work, we extend the notion of duality gap to proximal duality gap that is applicable to the general context of training GANs where Nash equilibria may not exist. We show theoretically that the proximal duality gap is capable of monitoring the convergence of GANs to a wider spectrum of equilibria that subsumes Nash equilibria. We also theoretically establish the relationship between the proximal duality gap and the divergence between the real and generated data distributions for different GAN formulations. Our results provide new insights into the nature of GAN convergence. Finally, we validate experimentally the usefulness of proximal duality gap for monitoring and influencing GAN training.

1. Introduction

Generative modeling is an important machine learning paradigm, aiming to learn data distributions. The ability to parametrically model the true underlying distribution of real-world data from a given empirical distribution brings with it the power to generate new and unseen instances. Generative adversarial network (GAN) is perhaps the most popular and successful of innovations for learning data distributions. A GAN formulates the generative modeling problem as a zero-sum game between two agents - a Discriminator (D) and a Generator (G). The discriminator aims to differentiate the fake samples produced by the generator from samples belonging to the true data distribution. On the other hand, the generator seeks to fool the discriminator by learning a mapping from an input noise space to the data space. The generator can also be viewed as performing adversarial attacks on the discriminator, exploiting the information leak through the discriminator and learning the real data distribution as the game proceeds to an equilibrium.

Formally the GAN game is defined as :

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

where the generator (parametrized by θ_g) and discriminator (parametrized by θ_d) are neural networks and V is the objective function that the agents seek to optimize. Different GAN formulations yield different expressions for V, each minimizing a unique divergence between the real and generated data distributions. The classic GAN formulation (Goodfellow et al., 2014) minimizes the JS divergence and is defined by :

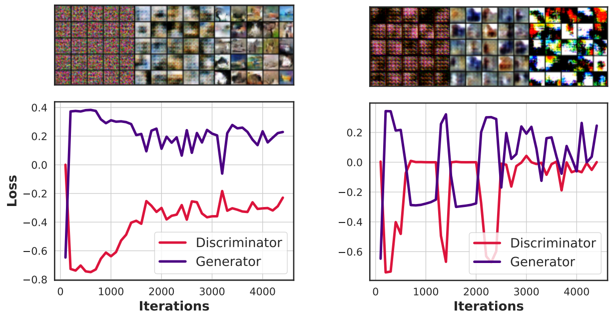
$$V = \mathbb{E}_{\mathbf{x}\sim P_r}[\log(D(\mathbf{x}))] + \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[\log(1 - D(\mathbf{x}))]$$

where P_r denotes the real data distribution and P_{θ_g} denotes the generated data distribution.

In any learning problem, the trajectory of the loss functions should indicate the goodness of the trained model. However, such intuitive inferences cannot be drawn from the loss curves of a GAN. This is because classical training of a GAN involves alternate gradient descent optimization of the objective function w.r.t the individual agents. Each optimization step of an agent alters its adversary's loss surface, resulting in non-intuitive loss curves for both the agents over time. Figure 1 shows discriminator and generator loss curves for a GAN when it (a) converges and (b) diverges. Ideally, losses should decrease during model convergence and increase during divergence. However, we observe a diminishing generator loss and an increasing discriminator loss when the GAN converges. When it diverges, there is an interplay between both the losses. These loss curves do not

[1]Department of Computer Science, Indian Institute of Technology, Ropar, India. Correspondence to: Sahil Sidheekh <2017csb1104@iitrpr.ac.in>, Aroof Aimen <2018csz0001@iitrpr.ac.in>, Narayanan C Krishnan <ckn@iitrpr.ac.in>.

Proceedings of the 37^{th} International Conference on Machine Learning, PMLR 108, 2021.



(a) Convergence (b) Divergence

Figure 1. Loss curves throughout the training progress of WGAN on the CIFAR-10 dataset.

give any insight into the gradual improvement or degradation in the GAN's performance. Thus, monitoring GANs often requires a subjective evaluation of the generated output. As a result, an exhaustive search over the architecture and hyperparameter space to find the delicate balance demanded by the GAN game becomes infeasible. This increases the complexity of GAN training that is already challenging due to the instabilities posed by the min-max gradient optimization. Objective measures capable of quantifying GAN training progress can reduce the training complexity.

Duality Gap (DG) (Grnarova et al., 2019) for GANs, motivated by principles of game theory, is a recently proposed objective measure for monitoring GAN training. The DG quantifies a GAN configuration's goodness in terms of the agents' ability to deviate from it in search of better optima. When a GAN converges to a Nash equilibrium, no agent can unilaterally deviate to find a better optima, and hence the DG would be zero. The ability to quantify convergence as well as the domain agnostic nature that requires no pre-trained models nor labeled data, makes DG a potentially powerful tool to monitor GAN training.

However, DG relies on the notion that GANs converge to Nash equilibria. On the contrary, recent studies (Farnia & Ozdaglar, 2020; Berard et al., 2019) suggest that a Nash equilibrium need not always exist for a GAN, especially when trained under regularized environments that most modern GAN formulations employ. GANs can converge to stable stationary points that are not Nash equilibria, all the while producing realistic data samples with high fidelity. This weakens the foundation upon which the notion of DG as a performance monitoring tool for GANs is built, eliciting the following questions: Can GANs capture the real data distribution even at non-Nash critical points? If so, would the DG at such stationary points be zero? If not, how do we monitor the GAN training in such situations?

In this work, we study the above questions by introducing the notion of proximal duality gap for GANs that is generalizable to scenarios where Nash equilibria may not exist. Our work is motivated by the notion of proximal equilibria for GANs that serves as a general concept for characterizing GAN optimality (Farnia & Ozdaglar, 2020). We define the DG in terms of the agents ability to optimize the proximal GAN objective (see Eq. 9) and call it as the Proximal Duality Gap (DG^{λ}). A proximal equilibrium for the GAN game (Eq. 1) is a Nash equilibrium w.r.t the proximal objective. Thus, whenever the GAN game attains a proximal equilibrium, DG^{λ} will tend to zero, indicating model convergence. As all Nash equilibria form a subset of proximal equilibria, DG^{λ} serves as a generic and robust measure that can quantify GAN convergence in the wild.

Overall, we make the following contributions:

- We present an acute limitation of DG for monitoring GAN training.
- We propose a theoretically grounded and robust extension - DG^{λ}, that overcomes this limitation and is also applicable to the broader context, when the GANs converge to a non-Nash equilibrium.
- Using DG^{λ}, we derive insights into the nature of GAN convergence. Specifically, we study the relationship between the quality of the learned data distribution and the game equilibria. We show that for various GANs, a configuration (θ_d, θ_g) where $P_{\theta_g} = P_r$ corresponds to a Stackelberg equilibrium.
- We demonstrate through experiments, the proficiency of DG^{λ} for monitoring and influencing GAN training.

2. Related Work

Motivated by the non-inferrable nature of GAN loss curves, developing extrinsic measures to monitor GAN training has emerged as an active research area (Borji, 2018; Lucic et al., 2018; Olsson et al., 2018). Existing measures such as average log-likelihood (Goodfellow et al., 2014; Theis et al., 2016), Inception Score (IS) (Salimans et al., 2016), Frechet Inception Distance (FID) (Heusel et al., 2017) evaluate the output of the GANs, but require pre-trained models. Further, these measures, including the more recent ones such as precision and recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019), density and coverage (Tolstikhin et al., 2017; Naeem et al., 2020) do not monitor the training progress nor characterize the equilibria of the GAN game.

Duality Gap (DG) (Grnarova et al., 2019) is a recently proposed domain agnostic and computationally feasible metric for monitoring and evaluating GAN training. DG is zero when GAN converges to a Nash equilibrium making it an attractive metric for objectively monitoring the GAN training. However, DG has a fundamental limitation with its estimation process due to vanishing gradients. Adding perturbations to the GAN configuration before estimating the duality gap (perturbed duality gap) helps to overcome the vanishing gradients (Sidheekh et al., 2020). But both these approaches assume the convergence of GANs to Nash equilibrium, which may not always be the case especially for high dimensional datasets (as we demonstrate in the next section). Thus, limiting the applicability of DG and perturbed DG for monitoring the GAN training.

3. Background

3.1. A Brief Overview of GAN formulations

Classic GAN : The min-max objective in the classic GAN (Goodfellow et al., 2014) formulation is :

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

where the probabilistic discriminator D outputs the likelihood of the input data point belonging to the real data distribution. The discriminator's objective is to maximize the log-likelihood to learn the conditional probability $P(y|\mathbf{x})$, where $y = 0$ and 1 indicate a fake and real data point respectively. Minimizing the above objective w.r.t the generator for the optimal discriminator is equivalent to minimizing the Jenson Shannon divergence (JSD) between P_{θ_g} and P_r.

F-GAN: F-GAN (Nowozin et al., 2016) is the generalization of the classic GAN to minimize arbitrary f-divergences by incorporating an extension of the variational divergence estimation framework (Nguyen et al., 2010). For a convex, lower semi-continuous function $f : \mathbb{R}_+ \to \mathbb{R}$ that satisfies $f(1) = 0$, the $f-$divergence between distributions P, and Q is,

$$D_f(P||Q) = \int p(x) f\left(\frac{q(x)}{p(x)}\right) dx \tag{3}$$

The F-GAN objective that minimizes the $f-$ divergence (D_f) between P_{θ_g} and P_r is defined as

$$V_f = \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[f^*(D(\mathbf{x}))] \tag{4}$$

where $f^*(x) = \sup_{t\in \text{Dom}(f)} \{xt - f(t)\}$ is the Fenchel-conjugate of f. The classic GAN is a special case of F-GAN when $f(t) = t\log(t) - (t+1)\log\left(\frac{t+1}{2}\right)$.

Wasserstein GAN (WGAN): WGAN formulates the GAN game as a minimization of the optimal transport cost, the Wasserstein distance, between P_r and P_{θ_g}, a more efficient cost function to learn data distributions having support on low dimensional manifolds (Arjovsky et al., 2017). Specifically, the Kantorovich-Rubinstein duality is used to arrive at the Wasserstein-1 (Earth Movers) distance between the distributions defined as: $\sup_{||D||_{L\leq 1}} \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[D(\mathbf{x})]$, where the supremum is over all 1-Lipschitz discriminators. In practice, the Lipschitz constraint is enforced through weight clipping resulting in the following WGAN objective:

$$V_{w_1} = \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[D(\mathbf{x})] \tag{5}$$

The WGAN formulation is also extended to a general transport cost (Farnia & Tse, 2018) $c(\mathbf{x}, \mathbf{y})$ by constraining the discriminators to be c-concave as

$$V_w = \mathbb{E}_{x\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{x\sim P_{\theta_g}}[D^c(\mathbf{x})], \tag{6}$$

where D^c is the c-transform of the discriminator D, i.e.,

$$D^c(\mathbf{x}) = \sup_{\mathbf{y}} \{D(\mathbf{y}) - c(\mathbf{x},\mathbf{y})\} \tag{7}$$

and the Wasserstein distance between P_{θ_g} and P_r is :

$$W_c(P_{\theta_g}||P_r) = \sup_{D-c-concave} \mathbb{E}_{\mathbf{x}\sim P_r}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{x}\sim P_{\theta_g}}[D^c(\mathbf{x})] \tag{8}$$

Despite the added stability of Wasserstein distance over other divergences, training GANs remains an arduous task. This has motivated efforts towards understanding the nature of GAN convergence.

3.2. Understanding GAN convergence

Classical Notion of GAN Equilibrium

Traditionally, the GAN game was expected to converge to a pure Nash equilibrium, a configuration (θ_d^*, θ_g^*) that is optimal for both the players i.e.

$$\max_{\theta_d} V(D_{\theta_d}, G_{\theta_g^*}) = \min_{\theta_g} V(D_{\theta_d^*}, G_{\theta_g}) = V(D_{\theta_d^*}, G_{\theta_g^*})$$

A GAN having unbounded capacity learns the true data distribution at such a solution (Goodfellow et al., 2014). However, a pure Nash equilibrium need not always exist for a zero sum game (Nash, 1950). Only an extended notion - the mixed strategy Nash equilibrium (MNE) is guaranteed to exist. Recent GAN formulations explicitly seek the MNE (Arora et al., 2017; Hsieh et al., 2019). As a mixed strategy gives a distribution over the model parameters, the stationary point to which the GAN converges need not be individually optimal for both the players.

GANs Need Not Converge to Nash Equilibria

GAN convergence has also been well studied from an optimization perspective using the notion of stability (Daskalakis et al., 2017; Fiez et al., 2019; Nouiehed et al., 2019; Zhang et al., 2019b; Mazumdar et al., 2019; Mokhtari et al., 2020; Lin et al., 2020). As GAN formulations are non-convex, only local surrogates of equilibria may be attainable while employing gradient based optimization. A (differential) local Nash equilibrium (LNE) (θ_d^*, θ_g^*) satisfies two properties: 1) $\nabla_{\theta_d} V(\theta_d^*, \theta_g^*) = \nabla_{\theta_g} V(\theta_d^*, \theta_g^*) = 0$

and 2) $\nabla^2_{\theta_d} V(\theta_d^*, \theta_g^*) \prec 0$, $\nabla^2_{\theta_g} V(\theta_d^*, \theta_g^*) \succ 0$. However, recent literature (Mazumdar & Ratliff, 2018; Adolphs et al., 2019) suggests the existence of many stable attractors that are not LNE for the alternate gradient descent optimization, nor produce realistic data and proposes methods to escape these attractors. (Fiez et al., 2019; Jin et al., 2020) study the gradient dynamics of sequential games and establish that their only stable attractors are Stackelberg equilibria. However, the relationship between the learned data distribution and the game configuration has not been well studied. Further, recent empirical studies (Berard et al., 2019; Farnia & Ozdaglar, 2020) also suggest that *GANs need not attain an LNE to produce realistic data*.

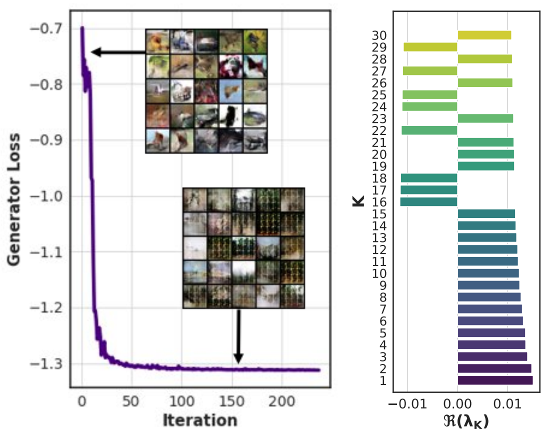


Figure 2. The high fidelity images outputted by a converged GAN deteriorates on optimizing only w.r.t the generator while attaining a lower loss (left), indicating that the GAN has not converged to a Nash equilibrium; confirmed by the positive and negative eigenvalues of the Hessian (right).

We verify the above hypothesis by training a spectral normalized GAN (SNGAN) on the CIFAR-10 dataset for 100 epochs ensuring that the models have converged producing high fidelity samples. As a local Nash equilibrium is locally optimal for the agents, optimizing the objective function w.r.t any individual player should not facilitate departure of the player from such a point. However, as demonstrated in Figure 2, the generator deviates from the attained equilibrium on further optimizing the objective function w.r.t only the generator. While the generator attains lower cost, there is a clear deterioration in the generated samples' quality. This implies that the GAN has not converged to a LNE. We further strengthen the claim by verifying the top-K eigenvalues (λ_K) (by magnitude) of the Hessian of the objective w.r.t the generator's parameters. However, as shown in Figure 2, the Hessian has both positive as well as negative eigen values violating the second property of an LNE, thus emphasizing that GANs can converge to non-Nash attractors, all the while producing high fidelity samples. Similar results on GANs trained for MNIST and CELEB-A datasets are discussed in the supplementary material.

Proximal Equilibria for GANs

Majority of studies attempting to characterize the equilibria for GANs assume unbounded capacity for the models (realizable setting) (Goodfellow et al., 2014; Arora et al., 2017; Hsieh et al., 2019). However, in practice, most GAN architectures (Brock et al., 2019; Zhang et al., 2019a; Miyato et al., 2018; Gulrajani et al., 2017; Arjovsky et al., 2017; Radford et al., 2016) employ normalization and regularization to achieve state of the art performance. A recent study (Farnia & Ozdaglar, 2020) on GAN convergence under the non-realizable setting proposes a more generic notion of equilibira - the Proximal Equilibria (PE). This notion of equilibira is derived from a sequential game-play perspective for a GAN, for which a Stackelberg equilibrium is guaranteed to exist under mild continuity assumptions (Jin et al., 2020; Fiez et al., 2019). Farnia and Ozdaglar define a proximal operator over the original GAN objective, allowing the discriminator to be optimal in a neighbourhood (controlled by λ),

$$V^{\lambda}(D_{\theta_d}, G_{\theta_g}) = \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) - \lambda ||D_{\tilde{\theta_d}} - D_{\theta_d}||^2 \tag{9}$$

A proximal equilibrium is defined as the Nash equilibrium for the objective V^{λ}, which is guaranteed to exist (Farnia & Ozdaglar, 2020). Formally, a configuration (θ_d^*, θ_g^*) of the GAN game (Eq 1) is called a $\lambda-$proximal equilibrium if and only if $\forall\ \theta_d, \theta_g$,

$$\begin{aligned} V(D_{\theta_d}, G_{\theta_g^*}) &\leq V(D_{\theta_d^*}, G_{\theta_g^*}) \\ &\leq \max_{\tilde{\theta_d} \in \Theta_D} V(D_{\tilde{\theta_d}}, G_{\theta_g}) - \lambda ||D_{\tilde{\theta_d}} - D_{\theta_d^*}||^2 \end{aligned} \tag{10}$$

As the extreme cases $\lambda \to \infty (0)$ recreate a Nash (Stackelberg) equilibrium, the $\lambda-$proximal equilibrium explores the spectrum of equilibria between the two and thus serves as a generic notion for GAN convergence. (Farnia & Ozdaglar, 2020) also suggest proximal training to explicitly enforce convergence to a proximal equilibrium. Our work, in contrast, is aimed towards evaluating GAN convergence and quantifying its goodness, irrespective of how it was trained.

4. Proximal Duality Gap

We first define the classical duality gap (Grnarova et al., 2019) for GANs before moving on to the proposed measure.

Definition 1. *Consider the GAN game presented in Eq.1. Then, for a configuration (θ_d, θ_g) of the game, the duality gap (DG) is defined as :*

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

At a pure Nash equilibrium (θ_d^*, θ_g^*), $DG(\theta_d^*, \theta_g^*) = 0$.

The DG has some interesting properties. First, it is lower bounded by the JSD between P_r and P_{θ_g}. Second, as DG is applicable to any GAN objective $V(D,G)$ and does not require pre-trained classifier nor labeled data, it is domain agnostic and potentially better equipped to monitor GAN training over other prevalent evaluation measures. However, as discussed previously, GANs can converge to non Nash attractors, where P_r and P_{θ_g} are aligned well. The DG at such equilibrium is not very well understood, limiting its practicality for monitoring GAN training.

We extend the notion of Duality Gap to the general context of training GANs where Nash equilibria need not be attainable, utilizing the proximal operator. We define the Proximal Duality Gap (DG^λ) as below.

Definition 2. *The proximal duality gap (DG^λ) at (θ_d, θ_g) for the GAN game presented in Eq.1 is defined as*

$$\begin{aligned} DG^\lambda(\theta_d, \theta_g) &= V_{D_w}(\theta_g) - V^\lambda_{G_w}(\theta_d), \text{ where} \\ V_{D_w}(\theta_g) &= \max_{\theta'_d \in \Theta_D} V(D_{\theta'_d}, G_{\theta_g}) \\ V^\lambda_{G_w}(\theta_d) &= \min_{\theta'_g \in \Theta_G} V^\lambda(D_{\theta_d}, G_{\theta'_g}) \end{aligned}$$

The terms D_w(or G_w) indicate the worst adversary that the generator (or discriminator) might face. Note that $\max_{\theta'_d \in \Theta_D} V^\lambda(D_{\theta'_d}, G_{\theta_g}) = \max_{\theta'_d \in \Theta_D} V(D_{\theta'_d}, G_{\theta_g})$. Thus, for a GAN configuration attained using V, the DG^λ measures the ability of the agents to deviate from it w.r.t the proximal objective V^λ.

Remark. *For all proximal equilibria (θ^*_d, θ^*_g) of the GAN game defined by Eq. 1, $V_{D_w}(\theta^*_g) = V^\lambda_{G_w}(\theta^*_d) = V^\lambda(\theta^*_d, \theta^*_g)$, thus $DG^\lambda(\theta^*_d, \theta^*_g) = 0$.*

This remark directly follows from the definition of proximal equilibria (Eq. 10). Thus DG^λ tending to zero implies that the GAN game has converged to a proximal equilibrium. However, as GANs are used for learning data distributions, it is important that measures to quantify GAN convergence should also give insights into the nature of the learned data distribution. Thus, to establish the applicability of DG^λ, we study how it relates to the divergence between the real and generated data distributions for various GAN formulations.

4.1. Theoretical Analysis

The definition of DG^λ has two terms - V_{D_w} and $V^\lambda_{G_w}$. We first theoretically establish the relationship between V_{D_w} and the divergences (DIV) used in various GAN formulations - the JS divergence ($JSD(P_{\theta_g}||P_r)$) for classical GAN objective V_c, the Wasserstein distance ($W_c(P_{\theta_g}||P_r)$) for the WGAN objective V_w, and the $f-$divergence ($D_f(P_{\theta_g}||P_r)$) for the F-GAN objective V_f.

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*

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

Proof. Deferred to the supplementary material. □

Thus, $V_{D_w}(\theta_g)$ measures the quality of the generator G_{θ_g}. If θ_g is optimal such that P_{θ_g} covers the true data distribution P_r, then $V_{D_w}(\theta_g)$ achieves the minimum value. In case of a mismatch between P_{θ_g} and P_r, either due to insufficient support or poor sample quality, $V_{D_w}(\theta_g)$ will increase, thus making it a potentially useful metric in itself to monitor GAN training. However, V_{D_w} does not incorporate the ability of the discriminator to deviate from the current game configuration. Thus, it cannot identify if the game has converged to an equilibrium. Further, as the minimum value for V_{D_w} will vary depending upon the GAN formulation, it cannot serve as a domain agnostic measure for monitoring GAN training. DG^λ, on the other hand will always tend to zero on attaining a proximal equilibrium irrespective of the GAN formulation. Thus, in the next result we analyze the behavior of DG^λ against the three GAN formulations. Specifically, we show that the DG^λ is positive and lower bounded the divergence between P_{θ_g} and P_r.

Theorem 1. *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 lower bounded by the JSD, Wasserstein distance, and $f-$divergence between the real (P_r) and generated (P_{θ_g}) data distributions respectively. i.e.,*

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

Proof. Deferred to the supplementary material. □

This theorem non-trivially extends the prior result on the bound of DG only for classic GAN (Grnarova et al., 2019).

4.2. Implications

As DG^λ is lower bounded by the divergence between the real and generated data distributions, $DG^\lambda \to 0$ not only implies that the GAN has reached an equilibrium, but also that generated distribution is close to the real data distribution.

Corollary. *For GAN formulations defined by V_c, V_w or V_f, the generator learns the real data distribution at a proximal equilibrium (θ^*_d, θ^*_g), as $DG^\lambda(\theta^*_d, \theta^*_g) = 0$.*

This furthers the adeptness of proximal equilibria to serve as a general optimality notion for GANs. As a proximal equilibrium need not be a Nash equilibrium, it also facilitates the following interesting observation

Remark. *GANs can capture the real data distribution even at non Nash game configurations.*

Thus, the empirical observation (Berard et al., 2019) that GANs can produce realistic data samples having high fidelity despite converging to a non-Nash attractor of the gradient dynamics is theoretically justified.

On similar lines, a natural question that arises concerning the behaviour of DG^{λ} is whether proximal equilibria constitute an exhaustive notion of equilibria at which GANs can capture the real data distribution. Precisely, we ask the question: Does GAN converging to a solution such that $P_{\theta_g} \rightarrow P_r \implies DG^{\lambda} \rightarrow 0$? Our answer begins with the following proposition concerning the extreme case for DG^{λ} as $\lambda \rightarrow 0$.

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; $P_{\theta_g^*} = P_r \implies DG^{\lambda=0}(\theta_d^*, \theta_g^*) = 0$.*

Proof. Deferred to the supplementary material. □

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. Deferred to the supplementary material. □

Thus $DG^{\lambda=0}(\theta_d, \theta_g) \rightarrow 0$ whenever $P_{\theta_g} \rightarrow P_r$. The value of λ restricts the discriminator in the proximal objective (V^{λ}) to be optimal within a neighbourhood. As $\lambda \rightarrow 0$, V^{λ} considers the optimal discriminator over the entire parameter space (Θ_D). Thus $DG^{\lambda=0}(\theta_d, \theta_g) = 0$ implies that (θ_d, θ_g) is a Stackelberg equilibrium. As all proximal equilibria form a subset of Stackelberg equilibria, $DG^{\lambda=0}$ would thus be an ideal choice to monitor GAN convergence. However, as λ decreases, the complexity of computing V^{λ} increases rapidly and becomes infeasible as $\lambda \rightarrow 0$. Hence, it is only practical to check if a GAN configuration is a $\lambda(>0)-$proximal equilibrium. But can DG^{λ}, for a fixed value of $\lambda(>0)$ monitor convergence of GANs to all $\lambda^{'}-$proximal equilibria? To address this question, let us study the two cases - (i) $\lambda^{'} \geq \lambda$ and (ii) $\lambda^{'} < \lambda$ separately. The following theorem addresses case (i) utilizing the hierarchical property of proximal equilibria.

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

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

Proof. Deferred to Supplementary material □

$DG^{\lambda^{'}}(\theta_d, \theta_g) = 0$ is a sufficient condition for (θ_d, θ_g) being a $\lambda^{'}-$proximal equilibrium. Thus, it follows from theorem 3 that DG^{λ} is adept to monitor convergence of GANs to all $\lambda^{'}(\geq \lambda)-$proximal equilibria. However, when a GAN converges to a $\lambda^{'}(< \lambda)-$proximal equilibrium, DG^{λ} can be prone to error. The following theorem addresses this issue by upper bounding the difference between DG^{λ} and the divergence between real and generated data distributions.

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. Deferred to Supplementary material □

Corollary. *For a GAN configuration (θ_d^*, θ_g^*) such that $P_{\theta_g^*} = P_r, DG^{\lambda}(\theta_d^*, \theta_g^*) < \epsilon$*

Thus even when the GAN converges to a $\lambda^{'}(< \lambda)-$proximal equilibrium, the error that DG^{λ} can incur is bounded. Previously, the absence of such an upper bound as implied by theorem 4 for DG, meant that DG need not necessarily be close to zero when P_{θ_g} is close to P_r. DG^{λ} however, rules out this possibility and is thus a theoretically grounded and robust measure that can serve as a tool for monitoring convergence of GANs in the wild.

4.3. Estimating Proximal Duality Gap

Computing the true DG^{λ} for a GAN configuration is a hard task as it involves finding the optima of non convex functions. We approximate DG^{λ} by employing gradient descent to estimate V_{D_w} and $V_{G_w}^{\lambda}$. For a GAN game governed by V, estimating V_{D_w} involves optimizing $V(\theta_d, \theta_g)$ w.r.t θ_d using gradient descent. However, estimating $V_{G_w}^{\lambda}$ requires gradient computation over the proximal operator. As shown in (Farnia & Ozdaglar, 2020), for a GAN objective function V that is smooth w.r.t θ_d, the gradient of the proximal objective (V^{λ}) w.r.t θ_g can be obtained in terms of V as : $\nabla_{\theta_g} V^{\lambda}(D_{\theta_d}, G_{\theta_g}) = \nabla_{\theta_g} V(D_{\theta_d^*}, G_{\theta_g})$, where θ_d^* represents the optimal discriminator implied by the proximal objective. Thus, to estimate $V_{G_w}^{\lambda}$, at every iteration we use gradient descent to obtain θ_d^* for the corresponding θ_g and update θ_g to minimize $V(\theta_d^*, \theta_g)$. The algorithm for the overall estimation process and the associated computational

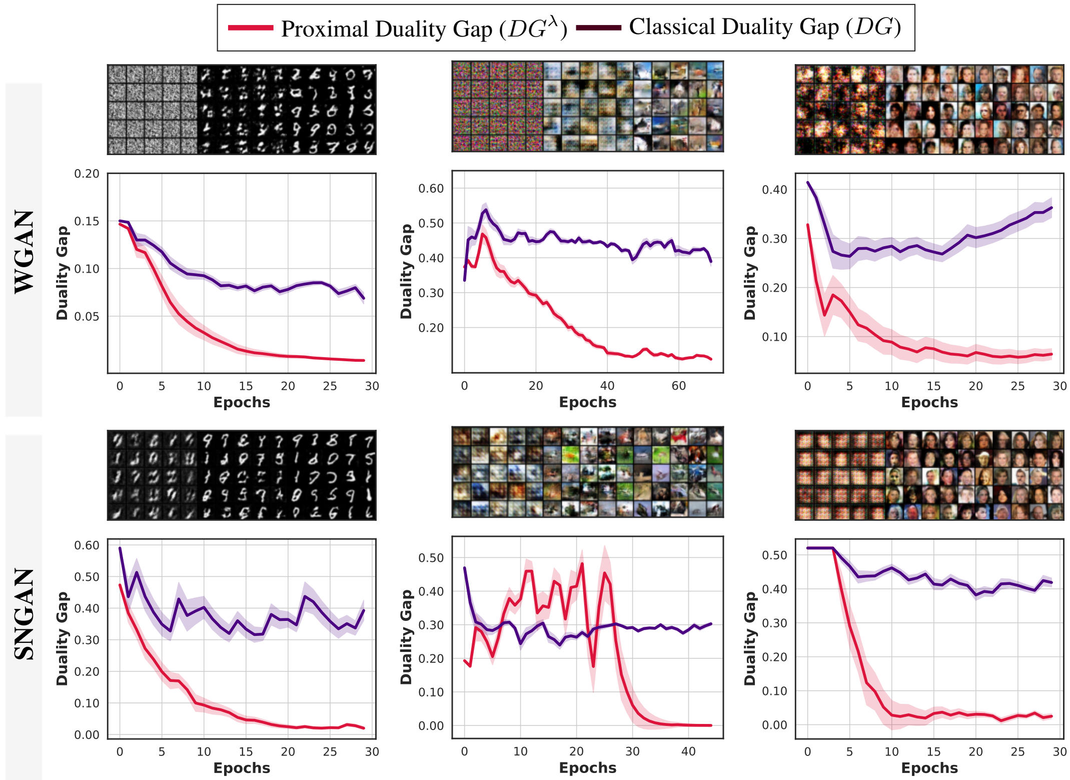


Figure 3. Monitoring Convergence of WGAN (Top Row) and SNGAN (Bottom Row) over the datasets MNIST (Col 1), CIFAR-10 (Col 2) and CELEB-A (Col 3) using duality gap. DG is not reflective of the GAN convergence. DG^λ, on the other hand is indicative of convergence and saturates close to zero. The shaded region indicates the standard deviation over 5 independent trials.

complexity are discussed in the supplementary material. To ensure that we obtain an unbiased estimate for DG^λ, following (Grnarova et al., 2019), we split the dataset into 3 disjoint sets - S_A, S_B and S_C. We train the GAN using S_A, we use 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.

5. Experimentation

To experimentally establish the proficiency of DG^λ, we consider a WGAN with weight-clipping (that optimizes V_w) (Arjovsky et al., 2017) and a Spectral Normalized GAN (SNGAN) (that optimizes V_c) (Miyato et al., 2018) over 3 datasets - MNIST (Deng, 2012), CIFAR-10 (Krizhevsky et al., 2014) and CELEB-A (Liu et al., 2015). For all the experiments, we use the 4-layer DCGAN (Radford et al., 2016) architecture for both the generator and the discriminator networks, and an Adam optimizer (Kingma & Ba, 2015) to train the models. To compute DG^λ, we use λ=0.1 and 20 optimization steps for approximating the proximal objective. We used the torchgan framework (Pal & Das, 2019) to train and evaluate all GAN models. Further implementation details for each experiment are provided in the supplementary material and the source code is publicly available [1].

	Pearson Correlation Coefficient (r)			
	$r_{DG,IS}$	$r_{DG^\lambda,IS}$	$r_{DG,FID}$	$r_{DG^\lambda,FID}$
MNIST	-0.752	$\mathbf{-0.892}$	0.845	$\mathbf{0.957}$
CIFAR-10	-0.368	$\mathbf{-0.854}$	0.213	$\mathbf{0.738}$
CELEB-A	-0.213	$\mathbf{-0.524}$	0.263	$\mathbf{0.699}$

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

Monitoring GAN training using DG$^\lambda$

Our first experiment aims to establish that DG^λ is better equipped over DG to monitor GAN convergence in practice. To this end, we train a WGAN and SNGAN over the 3 datasets till the models have converged. We compute DG and DG^λ throughout the training process, in addition to the (image) domain specific evaluation measures - IS and FID. Figure 3 demonstrates the training progress of each GAN, qualitatively through visualization of samples from

[1] https://github.com/proximal-dg/proximal_dg

the learned data distribution and quantitatively in terms of DG and DG^{λ}. The high fidelity of the learned data samples indicates that the models have converged. However, DG is not reflective of the training progress. This suggests that the GANs have not attained a Nash equilibrium - the behaviour of DG at non-Nash critical points is not well understood. DG^{λ}, on the other hand, captures the trend in the training progress, and eventually saturates close to zero. Thus, it is able to better characterize convergence. We also quantitatively validate the above observation by examining the correlation between DG and DG^{λ} against popular measures - IS and FID that quantify the quality of P_{θ_g}. As shown in Table 1, DG^{λ} has a higher positive correlation with FID and a higher negative correlation with IS as compared to DG. The duality gap is negatively correlated with IS because the latter increases as the GAN learns the real data distribution, whereas the former decreases. A larger (or smaller) IS (or FID) implies better fidelity of the learned data distribution. The higher correlation of DG^{λ} with IS and FID over the training process thus validates that DG^{λ} is adept to monitor not only the convergence of GANs to an equilibrium but also the goodness of P_{θ_g}.

Visualizing the effect of λ

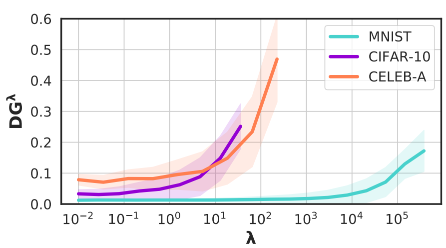


Figure 4. The behaviour of DG^{λ} for increasing λ

λ is a critical hyperparameter that determines the proficiency of DG^{λ}. We observed from the theoretical analysis that, while DG^{λ} is adept to monitor convergence of GANs to all $\lambda^{'}(\geq \lambda)$-proximal equilibria, it is prone to error as λ increases. As $\lambda \to \infty$, DG^{λ} becomes equivalent to DG. We thus experimentally study the behaviour of DG^{λ} for increasing values of λ. We compute DG^{λ} at the converged WGAN configurations (as shown in Fig 3) for each of the three datasets by varying λ in the range $[10^{-2}, 10^{6}]$. We observe (Figure 4) that for all the datasets, DG^{λ} remains close to zero and unaffected for λ in range $[10^{-2}, 1]$. Interestingly, for the MNIST dataset, DG^{λ} remains unaffected even for larger values of $\lambda(\approx 10^{4})$. This suggests that the WGAN configuration for MNIST is closer to a Nash equilibrium, also explaining why DG and DG^{λ} are closer for the same in Figure 3. As λ crosses a threshold (10^{1} for CELEB-A, CIFAR-10 and 10^{4} for MNIST), DG^{λ} increases sharply and behaves similar to DG. Thus, for a small value for λ (< 1) DG^{λ} is a robust tool for monitoring GAN convergence.

Influencing GAN training using DG$^{\lambda}$

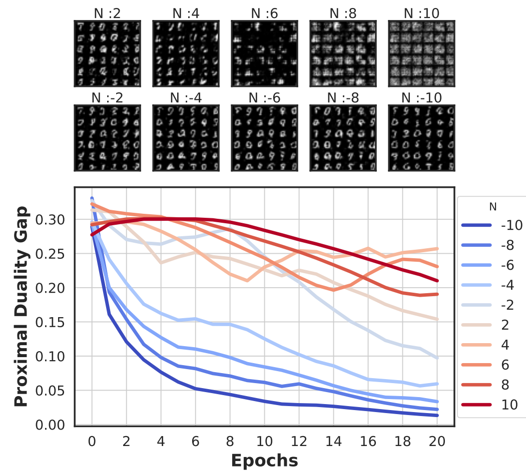


Figure 5. Tuning GAN Hyperparameter using DG^{λ}

A quantitative measure to monitor GAN training would enable easier tuning of hyperparameters. In this section, we explore DG^{λ} as an effective tool for influencing GAN training. A decisive hyperparameter that governs the delicate balance of the GAN game and hence its convergence is the update ratio of the agents. Let us denote by N, the number of discriminator updates per generator update, where a negative value for N indicates a larger number of generator updates. We train WGAN over the MNIST dataset by performing a grid search over N in the range -10 to 10 and computing DG^{λ}. Figure 5 depicts the qualitative output at the end of 20 epochs and DG^{λ} across training for each value of N. We observe that as N increases, the quality of the generated data samples diminishes and the learned data distribution eventually diverges as $N \to 10$. Correspondingly, we observe that DG^{λ} is close to zero for lower values of N(colored blue) and increases with N(colored red), suggesting that the GAN diverges for larger values of N. DG^{λ} thus enables us to quantitatively identify the optimal range of values for the hyperparameters of a GAN.

6. Summary and Future Work

GANs have pushed the boundaries of learning complex data distributions. However, the non-intuitive nature of GAN loss curves makes training a challenging task. We propose Proximal Duality Gap (DG^{λ}) as a generic and quantitative tool to monitor GAN training and understand GAN convergence. DG^{λ} characterizes GAN convergence as the game attaining a $\lambda-$proximal equilibrium. It also helps derive insights into the nature of GAN convergence - a GAN learns the real data distribution if and only if it attains a Stackelberg equilibrium. The ability of DG^λ to objectively quantify GAN convergence makes it a useful measure to tune the hyperparameters of a GAN. A couple of open questions that can improve the utility of DG^λ, if addressed, include identifying an optimal λ and making the range of values for DG^λ invariant across GAN formulations. The characterization of GAN convergence through proximal duality gap opens up new avenues for effortless GAN training.

References

Adolphs, L., Daneshmand, H., Lucchi, A., and Hofmann, T. Local saddle point optimization: A curvature exploitation approach. In *Proceedings of Machine Learning Research*, 2019.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of Machine Learning Research*, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, 2017.

Berard, H., Gidel, G., Almahairi, A., Vincent, P., and Lacoste-Julien, S. A closer look at the optimization landscapes of generative adversarial networks. *Computing Research Repository (CoRR)*, 2019.

Borji, A. Pros and cons of GAN evaluation measures. *Computing Research Repository (CoRR)*, 2018.

Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *Computing Research Repository (CoRR)*, 2017.

Deng, L. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, November 2012.

Farnia, F. and Ozdaglar, A. Do gans always have nash equilibria? In *International Conference on Machine Learning*, 2020.

Farnia, F. and Tse, D. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, 2018.

Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *Computing Research Repository (CoRR)*, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

Grnarova, P., Levy, K. Y., Lucchi, A., Perraudin, N., Goodfellow, I., Hofmann, T., and Krause, A. A domain agnostic measure for monitoring and evaluating gans. In *Advances in Neural Information Processing Systems*, 2019.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, 2017.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. In *International Conference on Machine Learning*, 2019.

Jin, C., Netrapalli, P., and Jordan, M. I. What is local optimality in nonconvex-nonconcave minimax optimization? *International Conference on Machine Learning*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Krizhevsky, A., Nair, V., and Hinton, G. Cifar-10 (canadian institute for advanced research). 2014.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, 2019.

Lin, T., Jin, C., and Jordan, M. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, 2015.

Lucic, M., Kurach, K., Michalski, M., Gelly, S., and Bousquet, O. Are gans created equal? a large-scale study. In *Advances in Neural Information Processing Systems*, 2018.

Mazumdar, E. and Ratliff, L. J. On the convergence of competitive, multi-agent gradient-based learning. *Computing Research Repository (CoRR)*, 2018.

Mazumdar, E. V., Jordan, M. I., and Sastry, S. S. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *Computing Research Repository (CoRR)*, 2019.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

Mokhtari, A., Ozdaglar, A., and Pattathil, S. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, 2020.

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable fidelity and diversity metrics for generative models. *Computing Research Repository (CoRR)*, 2020.

Nash, J. F. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 1950.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 2010.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, 2019.

Nowozin, S., Cseke, B., and Tomioka, R. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, 2016.

Olsson, C., Bhupatiraju, S., Brown, T. B., Odena, A., and Goodfellow, I. J. Skill rating for generative models. *Computing Research Repository (CoRR)*, 2018.

Pal, A. and Das, A. Torchgan: A flexible framework for gan training and evaluation. *Computing Research Repository (CoRR)*, 2019.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Machine Learning*, 2016.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, 2018.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

Sidheekh, S., Aimen, A., Madan, V., and Krishnan, N. C. On duality gap as a measure for monitoring gan training. *Computing Research Repository (CoRR)*, 2020.

Theis, L., van den Oord, A., and Bethge, M. A note on the evaluation of generative models. In Bengio, Y. and LeCun, Y. (eds.), *International Conference on Learning Representations*, 2016.

Tolstikhin, I. O., Gelly, S., Bousquet, O., Simon-Gabriel, C.-J., and Schölkopf, B. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, 2017.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, 2019a.

Zhang, K., Yang, Z., and Basar, T. Policy optimization provably converges to nash equilibria in zero-sum linear quadratic games. In *Advances in Neural Information Processing Systems*, 2019b.

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:

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

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:

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

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 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.*

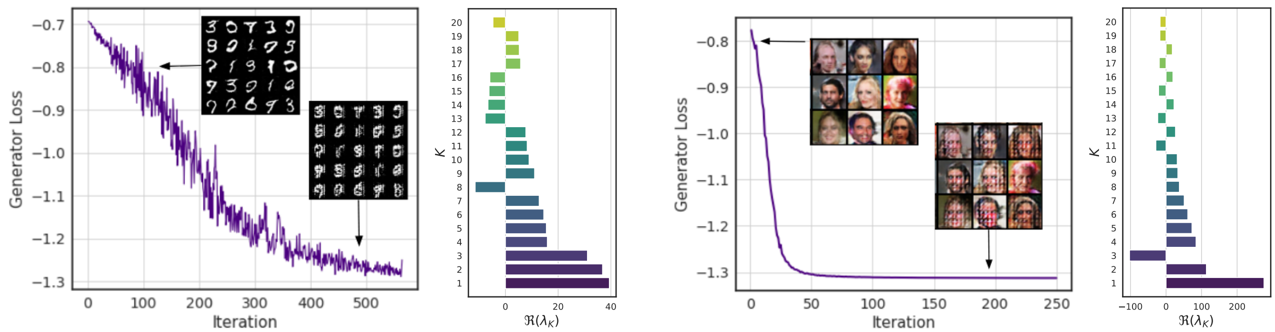
Proof. A configuration (θ_d^*, θ_g^*) of the GAN game (Eq 1) is called a Nash equilibrium if and only if $\forall\ \theta_d, \theta_g$,

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

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*

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

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,

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

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 $D_w(x) = D^c_w(x)$, 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,

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

$\square$

Theorem 1. *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 lower bounded by the JSD, Wasserstein distance, and $f-$divergence between the real (P_r) and generated (P_{θ_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)
$\implies V^\lambda(\theta_d,\theta_g) \leq 2JSD(P_{\theta_g}||P_r) - \log 4$

Thus, using a slight misuse of notation, we have

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

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

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

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

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

□

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

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

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

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

Part 3. *F-GAN*,

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

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

$\square$

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,

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

$\square$

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

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

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 $\forall\, \theta_d', \theta_g'$,

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

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

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

Thus, (θ_d, θ_g) is a $\lambda'-$proximal equilibrium $\implies (\theta_d, \theta_g)$ is also a λ_0-proximal equilibrium.
$\therefore DG^{\lambda=\lambda'}(\theta_d, \theta_g) = 0 \implies DG^{\lambda=\lambda_0}(\theta_d, \theta_g) = 0$ $\square$

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,*

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

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

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

□

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*(θ_d^t, θ_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 no. of gradient steps and becomes computationally infeasible as $\lambda \to 0$.

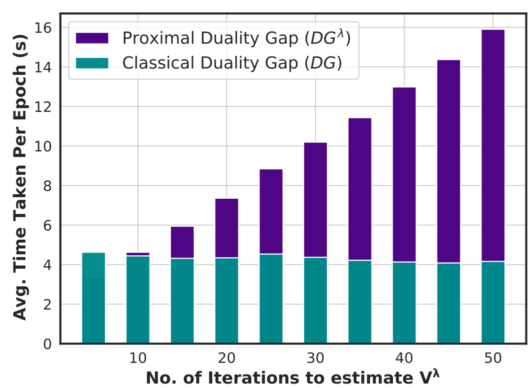


Figure 2. Computational Complexity of DG^{λ} over DG

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 $\lambda = 0.1$, $\approx$ 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

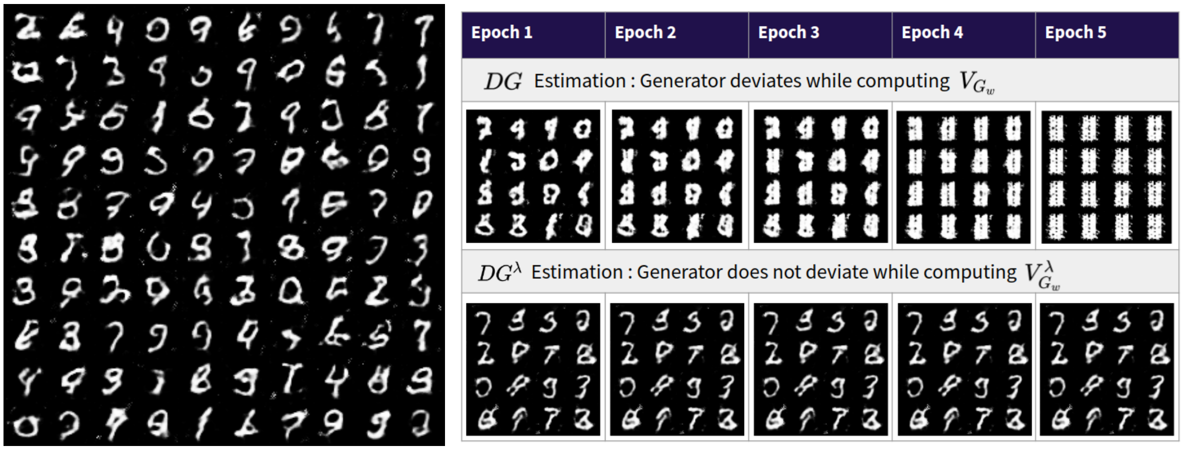
We used the 4-layer DCGAN architecture for both the generator and the discriminator networks in all the experiments. We used an Adam optimizer to train and evaluate all the models. To compute DG^{λ}, we used $\lambda = 0.1$ and 20 optimization steps for approximating the proximal objective. To enforce 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

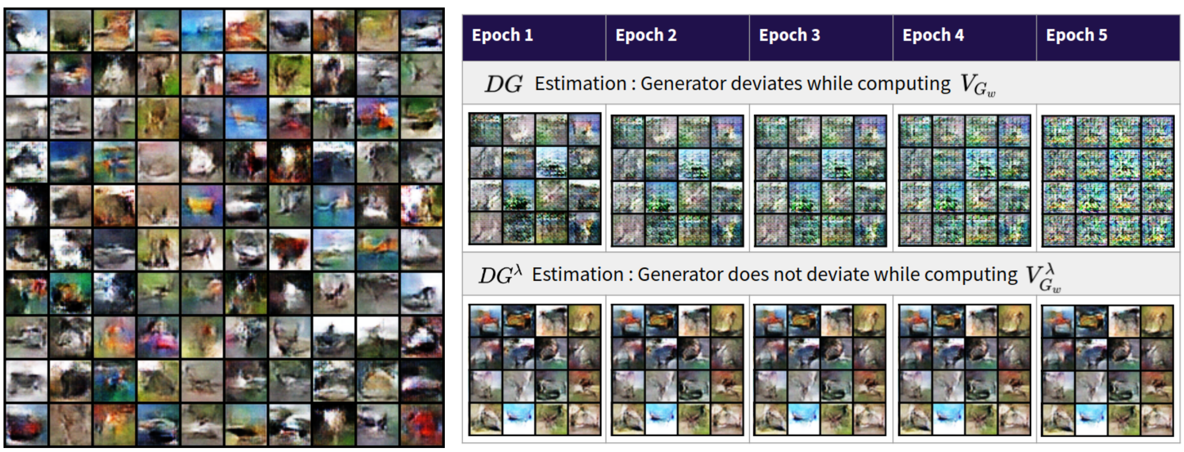
Table 2. Hyperparameter values for SNGAN experiments.

	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

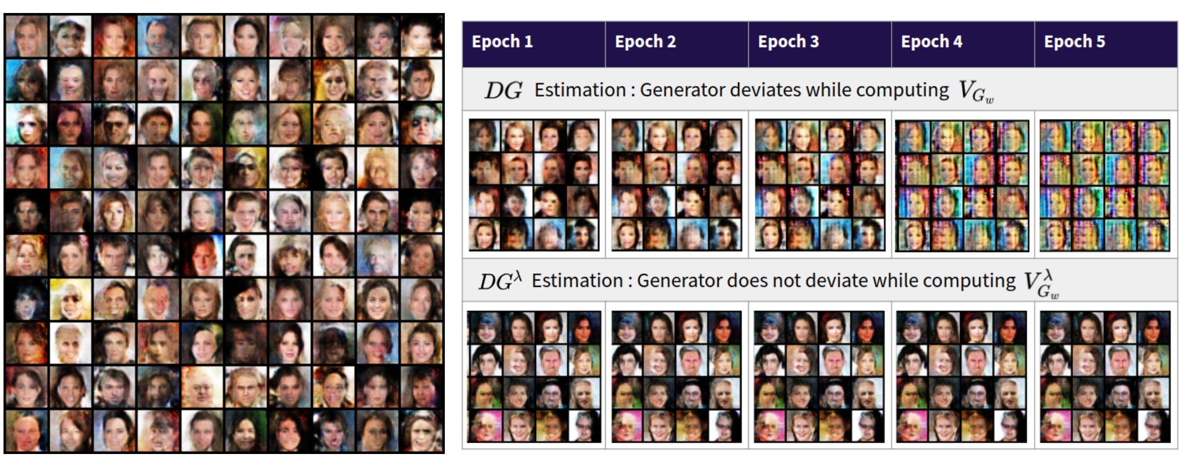
Table 3. Hyperparameter values for WGAN experiments.



(a) MNIST

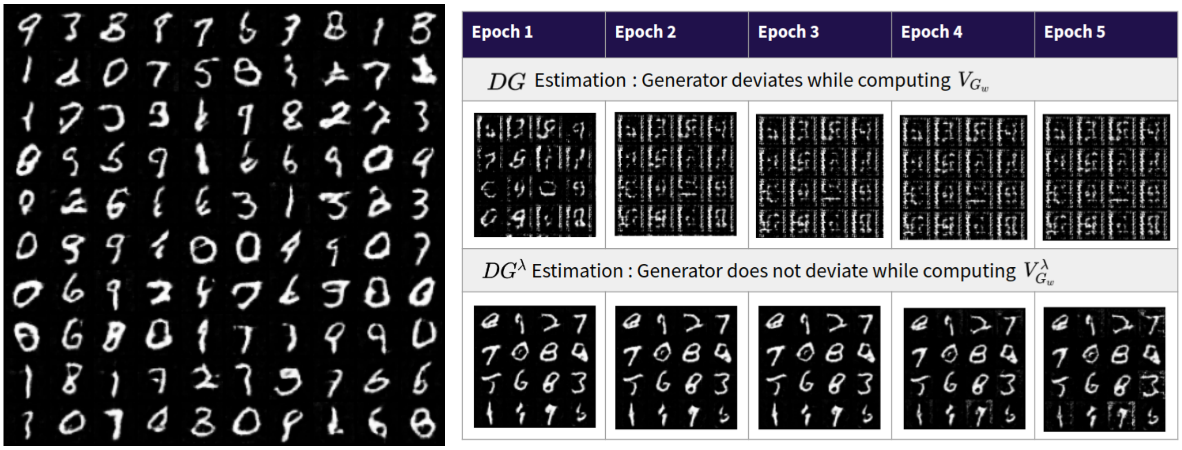


(b) CIFAR-10

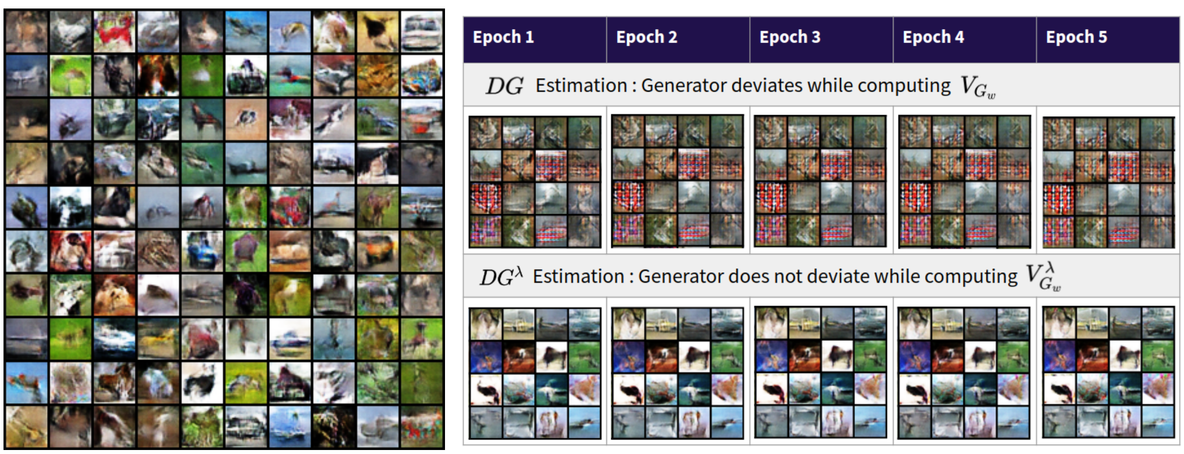


(c) CELEB-A

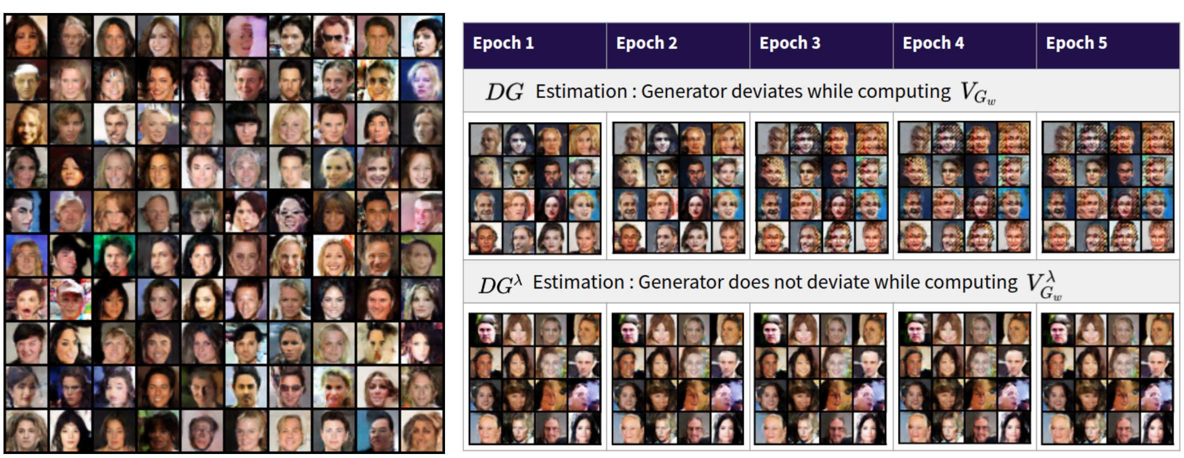
Figure 3. Visualizing the behaviour the generator while estimating DG and DG^{λ} for a converged Wasserstein GAN (WGAN) over the datasets- (a) MNIST (b) CIFAR-10 (c) CELEB-A. The GAN has converged, validated by the high fidelity of generated data samples(left grid). However, the generator deviates on 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 Normralized 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.