Do GANs always have Nash equilibria?

Farzan Farnia [1] **Asuman Ozdaglar** [1]

Abstract

Generative adversarial networks (GANs) represent a zero-sum game between two machine players, a generator and a discriminator, designed to learn the distribution of data. While GANs have achieved state-of-the-art performance in several benchmark learning tasks, GAN minimax optimization still poses great theoretical and empirical challenges. GANs trained using first-order optimization methods commonly fail to converge to a stable solution where the players cannot improve their objective, i.e., the Nash equilibrium of the underlying game. Such issues raise the question of the existence of Nash equilibria in GAN zero-sum games. In this work, we show through theoretical and numerical results that indeed GAN zero-sum games may have no Nash equilibria. To characterize an equilibrium notion applicable to GANs, we consider the equilibrium of a new zero-sum game with an objective function given by a proximal operator applied to the original objective, a solution we call the *proximal equilibrium*. Unlike the Nash equilibrium, the proximal equilibrium captures the sequential nature of GANs, in which the generator moves first followed by the discriminator. We prove that the optimal generative model in Wasserstein GAN problems provides a proximal equilibrium. Inspired by these results, we propose a new approach, which we call *proximal training*, for solving GAN problems. We perform several numerical experiments indicating the existence of proximal equilibria in GANs.

1. Introduction

Since their introduction in (Goodfellow et al., 2014), generative adversarial networks (GANs) have gained great success in many tasks of learning the distribution of observed samples. Unlike the traditional approaches to distribution learning, GANs view the learning problem as a zero-sum game between the following two players: 1) generator G aiming to generate real-like samples from a random noise input, 2) discriminator D trying to distinguish G's generated samples from real training data. This game is commonly formulated through a minimax optimization problem as follows:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} V(G,D). \tag{1}$$

Here, $\mathcal{G}$ and $\mathcal{D}$ are respectively the generator and discriminator function spaces, commonly chosen as two deep neural nets, and $V(G,D)$ denotes the minimax objective for generator G and discriminator D capturing how dissimilar G's produced samples and real training data are.

GAN optimization problems are commonly solved by alternating gradient methods, which under proper regularization have resulted in state-of-the-art generative models for various benchmark datasets (Goodfellow, 2016). However, GAN minimax optimization has introduced several theoretical and empirical challenges to the machine learning community. Training GANs is widely known as a challenging optimization task requiring an exhaustive hyper-parameter and architecture search and demonstrating an unstable behavior. While a few regularization schemes have led to empirical success in training GANs (Salimans et al., 2016; Arjovsky et al., 2017; Gulrajani et al., 2017; Miyato et al., 2018), still little is known about the conditions under which GAN minimax optimization can be successfully solved by first-order optimization methods.

To better understand the minimax optimization in GANs, one needs to first answer the following question: What is the proper notion of equilibrium in GAN zero-sum games? In other words, what should be the optimality criteria in the GAN's minimax optimization problem? A classical notion of equilibrium in the game theory literature is the *Nash equilibrium*, a state in which no player can improve its individual gain by choosing a different strategy. According to this definition, a Nash equilibrium (G^*, D^*) for the GAN minimax problem (1) must satisfy the following for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:

$$V(G^*, D) \le V(G^*, D^*) \le V(G, D^*). \tag{2}$$

[1]Laboratory for Information & Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. Correspondence to: Farzan Farnia <farnia@mit.edu>, Asuman Ozdaglar <asuman@mit.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

As a well-established result, for a generator G expressive enough to reproduce the distribution of observed samples, Nash equilibrium exists for that generator function producing the data distribution (Goodfellow, 2016). However, state-of-the-art GAN architectures (Gulrajani et al., 2017; Miyato et al., 2018; Zhang et al., 2018; Brock et al., 2018) commonly regularize the generator function through various means of regularization such as batch normalization or spectral regularization. Such regularization mechanisms do not allow the generator to match the empirical distribution of training data. Since the realizability assumption does not hold in such regularized GANs, Nash equilibria are not guaranteed to exist in their minimax optimization problem.

The above discussion motivates studying the equilibrium of GAN zero-sum games in *non-realizable settings* where the data distribution cannot be expressed by the regularized generator. Here, a natural question is whether a Nash equilibrium still exists in the non-realizable GAN problem. In this paper, we address this question and demonstrate through several theoretical and numerical results that:

Nash equilibrium may not always exist in GAN games.

We provide theoretical examples of well-known GAN formulations including the vanilla GAN (Goodfellow et al., 2014), Wasserstein GAN (WGAN) (Arjovsky et al., 2017), f-GAN (Nowozin et al., 2016), and 2-Wasserstein GAN (W2GAN) (Feizi et al., 2017) for which no local Nash equilibria exist. We further perform numerical experiments on widely-used GAN architectures which suggest that an empirically successful GAN training may converge to non-Nash equilibrium solutions.

Next, we focus on characterizing a new notion of equilibrium for GAN problems. To achieve this goal, we consider the Nash equilibrium of a new zero-sum game where the objective function is given by the following proximal operator applied to the minimax objective $V(G,D)$ with respect to a norm on discriminator functions:

$$V^{\text{prox}}(G,D) := \max_{\widetilde{D}\in\mathcal{D}} V(G,\widetilde{D}) - \left\|\widetilde{D}-D\right\|^2. \quad (3)$$

We refer to the Nash equilibrium of the new zero-sum game as the *proximal equilibrium*. Given the inherent sequential nature of GAN problems where the generator moves first followed by the discriminator, we consider a Stackelberg game for its representation and focus on the subgame perfect equilibrium (SPE) of the game as the right notion of equilibrium for such problems (Jin et al., 2019). We prove that the proximal equilibrium exists for Wasserstein GANs and provides an SPE for the GAN problem.

Inspired by these theoretical results, we propose a proximal approach for training GANs, which we call *proximal training*. In proximal training, we change the original minimax objective to the proximal objective in (3) and solve the following minimax problem via alternating gradient methods:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} V^{\text{prox}}(G,D). \quad (4)$$

In addition to preserving the SPE solution to the original GAN minimax problem, proximal training further enjoys the existence of Nash equilibrium solutions in its minimax objective. We discuss several numerical results supporting the proximal training approach and the role of proximal equilibrium solutions in Wasserstein and Lipschitz GAN problems. We can summarize the main contributions of this work as follows:

- Providing theoretical examples of standard GAN problems with no Nash equilibrium solutions,
- Introducing proximal equilibrium as a solution concept for GAN zero-sum games,
- Proving the existence of proximal equilibrium solutions for Wasserstein GANs,
- Proposing proximal training as a new training approach for GANs.

1.1. Related Work

Understanding minimax optimization in modern machine learning applications including GANs has been a subject of great interest in the machine learning literature. A large body of recent works (Daskalakis et al., 2017; Nouiehed et al., 2019; Mokhtari et al., 2019; Thekumparampil et al., 2019; Zhang et al., 2019; Mazumdar et al., 2019; Fiez et al., 2019; Wang et al., 2019; Lin et al., 2019) have analyzed the convergence properties of different optimization methods in solving various classes of minimax problems. The related references (Mertikopoulos et al., 2018; Bailey & Piliouras, 2018; Cheung & Piliouras, 2019; Flokas et al., 2019) study the complexities of reaching equilibrium solutions via first-order optimization methods in general minimax problems.

In a related work, Jin et al. (2019) propose a new notion of local optimality, called *local minimax*, designed for general sequential machine learning games. Compared to the notion of local minimax, the proximal equilibrium proposed in our work gives a notion of global optimality, which as we show directly applies to Wasserstein GANs. Jin et al. (2019) also provide examples of minimax problems with no Nash equilibrium solutions; however, the examples do not represent GAN minimax problems. The recent papers (Lin et al., 2019; Lei et al., 2019; Wang et al., 2020) have analyzed the convergence of different optimization methods to local minimax solutions. Also, Daskalakis & Panageas (2018) analyze the stable points of the gradient descent ascent (GDA) and optimistic GDA (Daskalakis et al., 2017) algorithms, proving that they can give strict supersets of

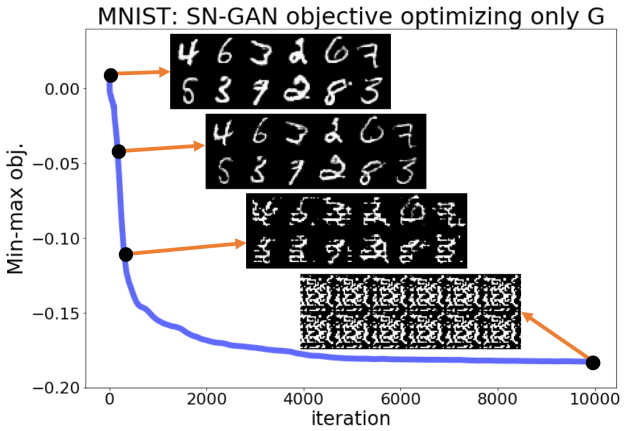

Figure 1. Optimizing the generator objective without changing the trained discriminator on the MNIST data. Both the SN-GAN objective and samples' quality were decreasing during the optimization.

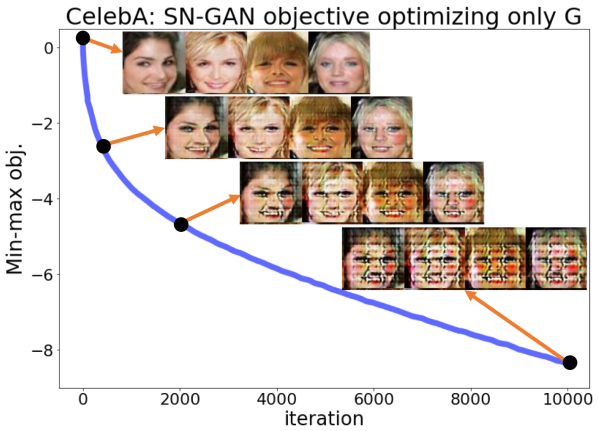

Figure 2. Repeating the experiment of Figure 1 on the CelebA dataset.

local saddle points. Regarding the stability of GANs, Nagarajan & Kolter (2017) prove that the GDA algorithm will be locally stable in GAN minimax problems with linear generator and discriminator functions. Feizi et al. (2017) show the GDA algorithm is globally stable for W2GANs with linear generator and quadratic discriminator functions.

Concerning the equilibrium properties of GANs, Berard et al. (2019) numerically demonstrate that state-of-the-art GAN architectures typically converge to stationary non-Nash equilibrium minimax points. Our numerical experiments in Section 2 provide additional numerical support for this paper's empirical results, and we further theoretically study the existence of Nash equilibrium solutions in GANs. References (Arjovsky & Bottou, 2017; Schäfer et al., 2019) perform complementary numerical experiments to study GANs' equilibrium solutions by fixing the trained generator and optimizing the discriminator. Fedus et al. (2017) empirically study the equilibrium of GAN problems regularized via the gradient penalty, reporting positive results on the stability of regularized GANs. However, our focus is on the existence of pure Nash equilibrium solutions.

Regarding the theoretical studies of equilibrium in GANs, Arora et al. (2017) study the equilibrium of GAN minimax games in realizable settings and also give an example of a simplified minimax GAN problem with no stable minimax solutions. Our work shows more realistic GAN problems with no Nash equilibrium solutions. The related papers (Arora et al., 2017; Hsieh et al., 2018) develop methods for finding mixed strategy Nash equilibria. On the other hand, our results focus on the pure strategies in non-realizable settings. Finally, developing GAN architectures with improved equilibrium and stability properties has been studied in several recent works (Metz et al., 2016; Mroueh et al., 2017; Berthelot et al., 2017; Heusel et al., 2017; Mescheder et al., 2017; Roth et al., 2017; Kodali et al., 2017; Daskalakis et al., 2017; Mescheder et al., 2018; Farnia & Tse, 2018; Sanjabi et al., 2018; Zhou et al., 2019; Taghvaei & Jalali, 2019).

2. Do GANs Empirically Converge to Nash Equilibria?

We empirically examined whether standard GAN architectures converge to Nash equilibrium solutions. In our numerical experiments, we applied three standard GAN architectures including the Wasserstein GAN with weight-clipping (WGAN-WC) (Arjovsky et al., 2017), the improved Wasserstein GAN with gradient penalty (WGAN-GP) (Gulrajani et al., 2017), and the spectrally-normalized vanilla GAN (SN-GAN) (Miyato et al., 2018) to standard MNIST (LeCun, 1998) and CelebA (Liu et al., 2015) datasets. We used the 4-layer convolutional neural network (CNN) architecture of the DC-GAN (Radford et al., 2015) and optimized the neural networks with the Adam (Kingma & Ba, 2014) or the RMSprop (Hinton et al., 2012) (only for WGAN-WC) algorithms. We followed all experimental details from the mentioned references.

We ran each of the GAN experiments for 200,000 generator iterations to reach $(G_{\boldsymbol{\theta}_{\text{final}}}, D_{\mathbf{w}_{\text{final}}})$ with $\boldsymbol{\theta}_{\text{final}}$ and $\mathbf{w}_{\text{final}}$ denoting the trained generator and discriminator parameters at the 200,000th iteration. We sought to examine whether $(G_{\boldsymbol{\theta}_{\text{final}}}, D_{\mathbf{w}_{\text{final}}})$ represents a Nash equilibrium. To do this, we fixed the trained discriminator $D_{\mathbf{w}_{\text{final}}}$ and kept optimizing the generator $G_{\boldsymbol{\theta}}$. Here we solved the following optimization problem initialized at $\boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}_{\text{final}}$ using the default first-order optimizer for 10,000 iterations:

$$\min_{\boldsymbol{\theta}} V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}_{\text{final}}}). \tag{5}$$

If the pair $(G_{\boldsymbol{\theta}_{\text{final}}}, D_{\mathbf{w}_{\text{final}}})$ was in fact a Nash equilibrium, it would provide a local saddle point for the minimax opti-

mization problem and the above optimization would not be able to find smaller objective values. Also, the image samples generated by $G_{\boldsymbol{\theta}}$ would improve or at least preserve their quality over this optimization, since $D_{\mathbf{w}_{\text{final}}}$ would achieve the same or better performance scores against other feasible generators.

However, we observed that none of the predicted outcomes hold for each of the six GAN experiments with three standard GAN architectures and two benchmark datasets. The optimization objective decreased rapidly from the beginning of the optimization, and the images sampled from the generator lost their quality over this optimization. Figures 1, 2 show the objective values for the SN-GAN experiments over the 10,000 steps of the described optimization. These figures also demonstrate the generated samples before and during the optimization, which shows the significant decrease in the quality of generated pictures. We defer the similar numerical results of the WGAN-WC and WGAN-GP experiments to the Appendix.

The above numerical results suggest that training GANs may not lead to local Nash equilibrium solutions in practice. After fixing the trained discriminator, the trained generator can be further optimized via a first-order optimization method to reach smaller values of the minimax objective. More importantly, this optimization not only does not improve the quality of the generated samples, but also completely disturbs the trained generator. As demonstrated in these experiments, simultaneous optimization of the two players is indeed necessary for proper convergence and stability in GAN optimization problems. The above experiments suggest that successfully-trained GANs need not converge to local Nash equilibria. In the upcoming sections, we review some standard GAN formulations and then show that there exist examples of GAN minimax problems with no Nash equilibrium solutions. Those theoretical results will further support the observations in the above experiments.

3. Existence of Nash Equilibria in GANs

3.1. Review of GAN formulations

Consider samples $\mathbf{x}_1, \dots, \mathbf{x}_n$ observed independently from underlying distribution $P_{\mathbf{X}}$. To find a generator function $G \in \mathcal{G}$ mapping a random noise input $\mathbf{Z}$ to the data distribution, Goodfellow et al. (2014) propose the following minimax problem called the *vanilla GAN*:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} \mathbb{E}\big[\log(D(\mathbf{X}))\big] + \mathbb{E}\big[\log(1 - D(G(\mathbf{Z})))\big]. \quad (6)$$

Here $\mathcal{G}$ and $\mathcal{D}$ represent the set of generator and discriminator functions, respectively. It can be seen that the above minimax problem with an unconstrained $\mathcal{D}$ containing all real-valued functions reduces to minimizing the Jensen-Shannon (JS) divergence between the data and generator's distributions. Nowozin et al. (2016) introduce f-GANs by extending the vanilla GAN to a general f-divergence. For a convex function $f : \mathbb{R} \to \mathbb{R}$, the f-divergence d_f is defined as $d_f(P, Q) := \int p(\mathbf{x}) f\big(\frac{q(\mathbf{x})}{p(\mathbf{x})}\big)\, \mathrm{d}\mathbf{x}$. Note that the JS-divergence is an f-divergence corresponding to $f_{\text{JSD}}(t) = t \log t - (t+1) \log \frac{t+1}{2}$. The f-GAN problem is formulated as:

$$\min_{G\in\mathcal{G}} \max_{D\in\mathcal{D}} \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[f^*\big(D(G(\mathbf{Z}))\big)\big], \quad (7)$$

where f^* denotes f's Fenchel-conjugate defined as $f^*(u) = \sup_t ut - f(t)$.

To resolve stability issues in training GANs, Arjovsky et al. (2017) formulate a GAN problem minimizing an optimal transport cost. Given a transportation cost function $c(\mathbf{x}, \mathbf{x}')$, the optimal transport cost W_c is defined as $W_c(P, Q) = \inf_{M\in\Pi(P,Q)} \mathbb{E}_M\big[c(\mathbf{X}, \mathbf{X}')\big]$. Here $\Pi(P, Q)$ denotes the set of all joint distributions on $(\mathbf{X}, \mathbf{X}')$ with $\mathbf{X}$, $\mathbf{X}'$ marginally distributed as P, Q, respectively. An important special case is the 1-Wasserstein (W_1) distance corresponding to $c(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|_2$. Formulating a minimax problem minimizing W_1-distance, Arjovsky et al. (2017) propose the following Wasserstein GAN (WGAN) problem where D is called 1-Lipschitz if for every $\mathbf{x}, \mathbf{x}'$ we have $D(\mathbf{x}) - D(\mathbf{x}') \le \|\mathbf{x} - \mathbf{x}'\|_2$:

$$\min_{G\in\mathcal{G}} \max_{D \text{ 1-Lipschitz}} \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[D(G(\mathbf{Z}))\big]. \quad (8)$$

The WGAN minimax problem can be generalized to other optimal transport costs with different cost functions. The generalization is as follows:

$$\min_{G\in\mathcal{G}} \max_{D \text{ c-concave}} \mathbb{E}\big[D(\mathbf{X})\big] - \mathbb{E}\big[D^c(G(\mathbf{Z}))\big], \quad (9)$$

where the c-transform is defined as $D^c(\mathbf{x}) = \sup_{\mathbf{x}'} D(\mathbf{x}') - c(\mathbf{x}, \mathbf{x}')$ and a function D is called c-concave if it is the c-transform of a valid function. In particular, the 2-Wasserstein GAN (W2GAN) problem (Feizi et al., 2017) considers a quadratic transportation cost $c(\mathbf{x}, \mathbf{x}') = \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|_2^2$.

3.2. Examples of GAN problems with no Nash equilibria

Consider a general GAN minimax problem (1) with a minimax objective $V(G, D)$. As discussed in the previous section, the optimal generator $G^* \in \mathcal{G}$ is defined to minimize the GAN's target divergence to the data distribution. The following proposition is a well-known result regarding the Nash equilibrium of the GAN game in realizable settings where there exists a generator $G \in \mathcal{G}$ producing the data distribution.

Proposition 1. *Assume that generator $G^* \in \mathcal{G}$ results in the distribution of data, i.e., we have $P_{G^*(\mathbf{Z})} = P_{\mathbf{X}}$. Then,*

for each of the GAN problems discussed in Section 3.1 there exists a constant discriminator function D_{constant} which together with G^* results in a Nash equilibrium for the GAN game, and hence satisfies the following for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$:

$$V(G^*, D) \leq V(G^*, D_{\text{constant}}) \leq V(G, D_{\text{constant}}).$$

Proof. This proposition is well-known for the vanilla GAN (Goodfellow, 2016). In the Appendix, we provide a proof for general f-GANs and Wasserstein GANs. □

The above proposition shows that in a realizable setting with a generator function generating the distribution of observed samples, a Nash equilibrium exists for that optimal generator. However, the realizability assumption in this proposition does not always hold in real GAN experiments. For example, in the GAN experiments discussed in Section 2, we observed that the divergence estimate never reached the zero value because of regularizing the generator function. Therefore, the Nash equilibrium described in Proposition 1 does not apply to the trained generator and discriminator in such GAN experiments.

Here, we address the question of the existence of Nash equilibrium solutions for non-realizable settings, where no generator $G \in \mathcal{G}$ can produce the data distribution. Do Nash equilibria always exist in non-realizable GAN zero-sum games? The following theorem shows that the answer is in general no. Note that $\sigma_{\max}(\cdot)$ in this theorem denotes the maximum singular value, i.e., the spectral norm.

Theorem 1. *Consider a GAN minimax problem for learning a normally distributed $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ with zero mean and scalar covariance matrix where $\sigma > 1$. In the GAN formulation, we use a linear generator function $G(\mathbf{z}) = \mathbf{W}\mathbf{z} + \mathbf{u}$ where the weight matrix $\mathbf{W}$ and vector $\mathbf{u}$ are regularized to satisfy $\sigma_{\max}(\mathbf{W}) \leq 1$ and $\|\mathbf{u}\|_2 \leq t$ for a constant $t > 0$. Suppose that the latent vector $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, I)$ has an isotropic Gaussian distribution. Then,*

- *For the f-GAN problem (7) corresponding to an f with non-decreasing $t^2 f''(t)$ over $t \in (0, +\infty)$ and an unconstrained discriminator D where the dimensions of $\mathbf{X}, \mathbf{Z}$ are equal, the f-GAN minimax problem has no Nash equilibrium solutions.*
- *For the W2GAN problem (9) with discriminator D trained over c-concave functions, where c is the quadratic cost, the W2GAN minimax problem has no Nash equilibrium solutions. Also, given a quadratic discriminator $D(\mathbf{x}) = \mathbf{x}^T A \mathbf{x} + \mathbf{b}^T \mathbf{x}$ parameterized by $A, \mathbf{b}$, the W2GAN problem has no **local** Nash equilibria.*
- *For the WGAN problem (8) with 1-dimensional X, Z and a discriminator D trained over 1-Lipschitz functions, the WGAN minimax problem has no Nash equilibria.*

Proof. We defer the proof to the Appendix. Note that the condition on the f-GAN holds for all f-GAN examples in (Nowozin et al., 2016) including the vanilla GAN. □

The above theorem shows that under the stated assumptions the GAN zero-sum game does not have Nash equilibrium solutions. Consequently, the optimal generator minimizing the distance to the data distribution does not provide a Nash equilibrium. While Theorem 1 provides examples of non-realizable GAN problems with no Nash equilibrium solutions, the following remark shows that non-reliability does not always imply the non-existence of Nash equilibria.

Remark 1. *Consider the same setting as in Theorem 1. However, unlike Theorem 1 suppose that $\sigma < 1$ and $\sigma_{\min}(\mathbf{W}) \geq 1$ where $\sigma_{\min}$ stands for the minimum singular value. Then, for the WGAN and W2GAN problems described in Theorem 1, the Wasserstein distance-minimizing generator results in a Nash equilibrium.*

Proof. We defer the proof to the Appendix. □

The above remark explains that the phenomenon shown in Theorem 1 does not always hold in non-realizable GAN settings. As a result, we need other notions of equilibrium which consistently explain optimality in GAN games.

4. Proximal Equilibria in GANs

4.1. Proximal equilibrium: A relaxation of Nash equilibrium

Since Theorem 1 shows that Nash equilibria are not guaranteed to exist for GANs, we consider a sequential Stackelberg competition to model the GAN zero-sum game in which the generator moves first followed by the discriminator. Note that the subgame perfect equilibrium (SPE) (G^*, D^*) of this Stackelberg game will satisfy:

$$G^* \in \operatorname*{argmin}_{G \in \mathcal{G}} \{\max_{D \in \mathcal{D}} V(G, D)\},$$
$$D^* \in \operatorname*{argmax}_{D \in \mathcal{D}} V(G^*, D). \tag{10}$$

Such an SPE solution, which we call Stackelberg equilibrium, will exist in the sequential GAN game under mild continuity assumptions (Jin et al., 2019). However, finding the above solution requires considering the maximized discriminator objective as the generator's cost function which will be computationally complex in general.

Here, we propose a new notion of equilibrium called proximal equilibrium which allows us to explore the spectrum between Nash and Stackelberg equilibria. In a proximal equilibrium, we allow the discriminator to further locally optimize itself in a norm-ball around the original discriminator.

This property is in fact consistent with the stability behavior observed for GANs trained by first-order optimization methods where the alternating first-order method stabilizes around a certain solution. To define proximal equilibria, we first define the following proximal objective for the original minimax objective $V(G, D)$:

$$V_\lambda^{\text{prox}}(G, D) := \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda \big\| \widetilde{D} - D \big\|^2. \quad (11)$$

The above definition represents the application of a proximal operator to $V(G, D)$, which further optimizes the original objective in the proximity of discriminator D. To keep the $\widetilde{D}$ function variable close to D, we penalize the distance between the two functions in the proximal optimization. Here the distance is measured according to norm function $\|\cdot\|$ on the discriminator space. We propose considering the Nash equilibria of the defined proximal objective $V_\lambda^{\text{prox}}(G, D)$ and define them as the proximal equilibria of $V(G, D)$.

Definition 1. *We call (G^*, D^*) a λ-proximal equilibrium for $V(G, D)$ if it represents a Nash equilibrium for $V_\lambda^{\text{prox}}(G, D)$, i.e. for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$*

$$V_\lambda^{\text{prox}}(G^*, D) \leq V_\lambda^{\text{prox}}(G^*, D^*) \leq V_\lambda^{\text{prox}}(G, D^*). \quad (12)$$

The next proposition provides necessary and sufficient conditions in terms of the original objective $V(G, D)$ for the proximal equilibrium solutions.

Proposition 2. *(G^*, D^*) is a λ-proximal equilibrium if and only if for every $G \in \mathcal{G}$ and $D \in \mathcal{D}$ we have*

$$V(G^*, D) \leq V(G^*, D^*) \leq \max_{\widetilde{D} \in \mathcal{D}} V(G, \widetilde{D}) - \lambda \big\| \widetilde{D} - D^* \big\|^2.$$

Proof. We defer the proof to the Appendix. □

Corollary 1. *Suppose that (G^*, D^*) is a λ-proximal equilibrium of $V(G, D)$ for $\lambda > 0$. Then, (G^*, D^*) provides a Stackelberg equilibrium for the minimax objective $V(G, D)$ and satisfies the equations in (10).*

The above corollary shows that every proximal equilibrium solution will provide a Stackelberg equilibrium for the GAN minimax problem, and therefore the generator G^* at a proximal equilibrium will minimize the maximum discriminator objective, i.e., the distance to the data distribution. The following result further shows that proximal equilibria satisfy a nested property and provide a hierarchy of equilibrium solutions for different λ values.

Proposition 3. *Define $\text{PE}_\lambda(V)$ to be the set of the λ-proximal equilibria for $V(G, D)$. Then, if $\lambda_1 \leq \lambda_2$,*

$$\text{PE}_{\lambda_2}(V) \subseteq \text{PE}_{\lambda_1}(V). \quad (13)$$

Proof. We defer the proof to the Appendix. □

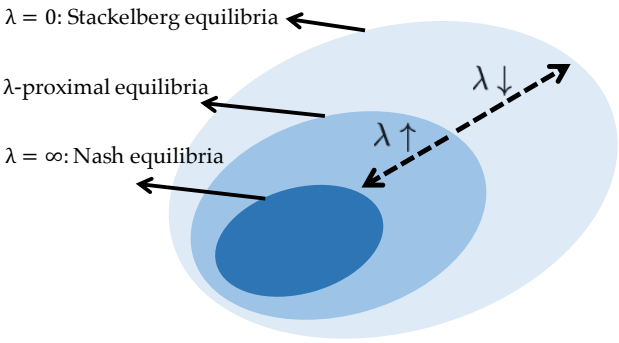

Figure 3. Proximal equilibria for different λ values with Nash ($\lambda = \infty$) and Stackelberg ($\lambda = 0$) equilibria as the two extremes.

Note that as λ approaches infinity, $V_\lambda^{\text{prox}}(G, D)$ tends to the original $V(G, D)$, implying that $\text{PE}_{\lambda=\infty}(V)$ is the set of $V(G, D)$'s Nash equilibria. In contrast, for $\lambda = 0$ the proximal objective becomes the worst-case objective $\max_{D \in \mathcal{D}} V(G, D)$. As a result, $\text{PE}_{\lambda=0}(V)$ reduces to the set of Stackelberg equilibria. Figure 3 illustrates the mentioned nested property of proximal equilibria for different λ values with the discussed extreme cases, i.e. Nash equilibria for $\lambda = \infty$ and Stackelberg equilibria for $\lambda = 0$.

Concerning the proximal optimization problem in (11), the following proposition shows that if the original minimax objective is a smooth function of the discriminator parameters, the proximal optimization can be solved efficiently and therefore one can compute the gradient of the proximal objective.

Proposition 4. *Consider the maximization problem in the definition of proximal objective (11) where generator $G_{\boldsymbol{\theta}}$ and discriminator $D_{\mathbf{w}}$ are parameterized by vectors $\boldsymbol{\theta}$, $\mathbf{w}$, respectively. Suppose that*

- *For the considered discriminator norm $\|\cdot\|$, $\|D_{\mathbf{w}} - D\|^2$ is η_1-strongly convex in $\mathbf{w}$ for any function D, i.e. for any $\mathbf{w}, \mathbf{w}', D$:*

$$\big\| \nabla_{\mathbf{w}} \|D_{\mathbf{w}} - D\|^2 - \nabla_{\mathbf{w}} \|D_{\mathbf{w}'} - D\|^2 \big\|_2 \geq \eta_1 \big\| \mathbf{w} - \mathbf{w}' \big\|_2,$$

- *For every $G_{\boldsymbol{\theta}}$, The GAN minimax objective $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}})$ is η_2-smooth in $\mathbf{w}$, i.e.*

$$\big\| \nabla_{\mathbf{w}} V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) - \nabla_{\mathbf{w}} V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) \big\|_2 \leq \eta_2 \|\mathbf{w} - \mathbf{w}'\|_2.$$

Under the above assumptions, if $\eta_2 < \lambda\eta_1$, the maximization objective in (11) is $\big(\lambda\eta_1 - \eta_2\big)$-strongly concave. Then, the maximization problem has a unique solution $\mathbf{w}^*$ and if $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}})$ is differentiable with respect to θ we have

$$\nabla_\theta V_\lambda^{\text{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}) = \nabla_\theta V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}^*}). \quad (14)$$

Proof. We defer the proof to the Appendix. □

The above proposition suggests that under the mentioned assumptions one can efficiently compute the optimal solution to the proximal maximization through a first-order optimization method. The assumptions require the smoothness of the GAN minimax objective with respect to the discriminator parameters, which can be imposed by applying norm-based regularization tools to neural network discriminators. Therefore, Proposition 4 shows that the complexity of solving the proximal optimization problem will decrease for a larger λ; however, for a sufficiently large λ the proximal objective may have no Nash equilibria, as shown in Theorem 1.

4.2. Proximal equilibria exist for Wasserstein GANs

As shown earlier, GAN minimax problems may not have any Nash equilibria in non-realizable cases. As a result, we should consider a different notion of equilibrium which is guaranteed to exist for GAN problems. As already discussed, Stackelberg equilibria defined in (10) will always exist for GANs and are special cases of λ-proximal equilibria for $\lambda = 0$. However, as implied by Proposition 4 a smaller λ means a greater optimization complexity for finding the equilibrium solution. Does there exist a positive $\lambda > 0$ for which the distance-minimizing generator provides a λ-proximal equilibrium? In this section, we show that such a positive λ in fact exists for Wasserstein GANs.

To define an appropriate proximal operator for Wasserstein GAN problems, we use a Sobolev semi-norm averaged over the underlying distribution of data $P_{\mathbf{X}}$. Here, we consider the following Sobolev norm function:

$$\|D\|_{\dot{H}^1} := \sqrt{\mathbb{E}_{P_{\mathbf{X}}}\left[\left\|\nabla_{\mathbf{x}} D(\mathbf{X})\right\|_2^2\right]}. \tag{15}$$

The above semi-norm is induced by the following semi-inner product and therefore leads to a semi-Hilbert space of functions:

$$\langle D_1, D_2\rangle_{\dot{H}^1} := \mathbb{E}_{P_{\mathbf{X}}}\left[\nabla D_1(\mathbf{X})^T \nabla D_2(\mathbf{X})\right]. \tag{16}$$

Throughout our discussion, we consider a parameterized set of generators $\mathcal{G} = \{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$. For a GAN minimax objective $V(G, D)$, we define $D^{\boldsymbol{\theta}}$ to be the optimal discriminator function for the parameterized generator $G_{\boldsymbol{\theta}}$:

$$D^{\boldsymbol{\theta}} := \operatorname*{argmax}_{D\in\mathcal{D}} V(G_{\boldsymbol{\theta}}, D). \tag{17}$$

The following theorem shows that the Wasserstein distance-minimizing generator in the 2-Wasserstein GAN (W2GAN) problem satisfies the conditions of a proximal equilibrium based on the Sobolev semi-norm in (15).

Theorem 2. *Consider the 2-Wasserstein GAN problem (9) with a quadratic cost $c(\mathbf{x}, \mathbf{x}') = \eta\|\mathbf{x} - \mathbf{x}'\|_2^2$. Suppose that the set of optimal discriminators $\{D^{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is convex. Then, $(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*})$ for the Wasserstein distance-minimizing generator $G_{\boldsymbol{\theta}^*} \in \mathcal{G}$ will provide a $\frac{1}{4\eta}$-proximal equilibrium with respect to the Sobolev norm in (15).*

Proof. We defer the proof to the Appendix. □

Algorithm 1 GAN Proximal Training

Input: data $\mathbf{x}_i$, size n

Initialize the parameters $\mathbf{w}^{(0)}, \boldsymbol{\theta}^{(0)}$

for k $= 0$ **to** MAX_ITER **do**

Initialize $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)}$

for t $= 0$ **to** T **do**

» $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k+1)} + \gamma_{k,t}\nabla_{\mathbf{w}}\{V(G_{\boldsymbol{\theta}^{(k)}}, D_{\mathbf{w}}) - \frac{\lambda}{2n}\sum_{i=1}^{n}\|\nabla_{\mathbf{x}} D_{\mathbf{w}}(\mathbf{x}_i) - \nabla_{\mathbf{x}} D_{\mathbf{w}^{(k)}}(\mathbf{x}_i)\|_2^2\}$

end for

» $\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma_k \nabla_{\boldsymbol{\theta}} V(G_{\boldsymbol{\theta}^{(k)}}, D_{\mathbf{w}^{(k+1)}})$.

end for

The above theorem shows that while, as demonstrated in Theorem 1, the W2GAN problem may have no local Nash equilibrium solutions, the proximal equilibrium exists for the W2GAN problem and holds at the Wasserstein-distance minimizing generator $G_{\boldsymbol{\theta}^*}$. The next theorem extends this result to the Wasserstein GAN (WGAN) problem minimizing the 1-Wasserstein distance.

Theorem 3. *Consider the WGAN problem (8) minimizing the first-order Wasserstein distance. For each $G_{\boldsymbol{\theta}}$, define $\alpha_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ to be the magnitude of the resulted optimal transport map from $\mathbf{X}$ to $G_{\boldsymbol{\theta}}(\mathbf{Z})$, i.e. $\mathbf{X} - \alpha_{\boldsymbol{\theta}}^2(\mathbf{X})\nabla D^{\boldsymbol{\theta}}(\mathbf{X})$ shares the same distribution with $G_{\boldsymbol{\theta}}(\mathbf{Z})$.[1] Given these definitions, assume that*

- *$\{\alpha_{\boldsymbol{\theta}}(\cdot)\nabla D^{\boldsymbol{\theta}}(\cdot) : \boldsymbol{\theta} \in \Theta\}$ is a convex set,*
- *for every $\mathbf{x}$ and $\boldsymbol{\theta}$, $\eta \leq \alpha_{\boldsymbol{\theta}}^2(\mathbf{x})$ holds for constant η.*

Then, $(G_{\boldsymbol{\theta}^*}, D^{\boldsymbol{\theta}^*})$ for the Wasserstein distance-minimizing generator function $G_{\boldsymbol{\theta}^*}$ gives an η-proximal equilibrium with respect to the Sobolev norm in (15).

Proof. We defer the proof to the Appendix. □

The above theorem shows that if the magnitude of optimal transport map is everywhere lower-bounded by λ, then the Wasserstein distance-minimizing generator in the WGAN problem yields a λ-proximal equilibrium.

[1] Note that as shown in the proof such a mapping $\alpha_{\boldsymbol{\theta}}$ exists under mild regularity assumptions.

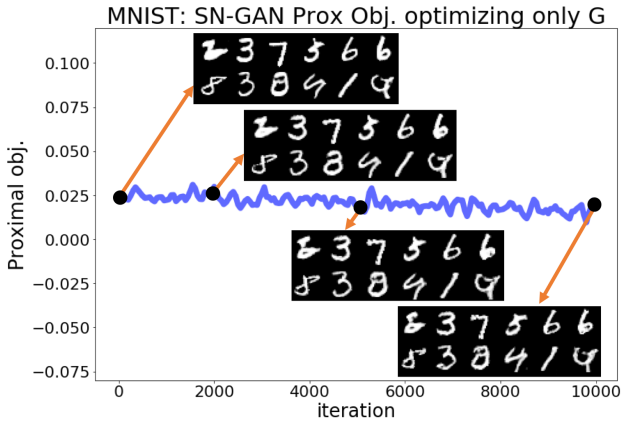

Figure 4. Optimizing the proximal objective over the generator with a fixed discriminator on MNIST data. The SN-GAN's objective and samples' quality were preserved during the optimization.

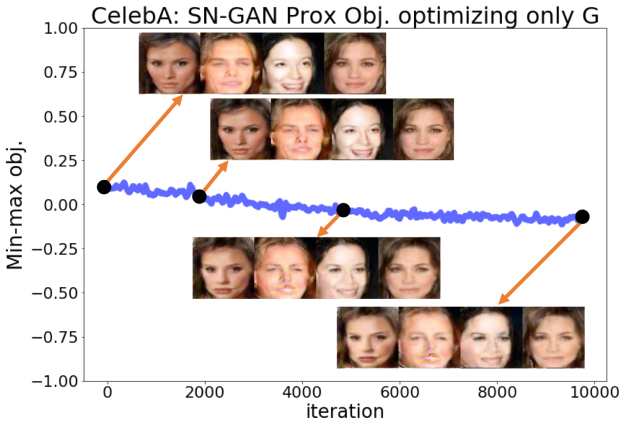

Figure 5. Repeating the experiment of Figure 4 for the CelebA dataset.

5. Proximal Training

As shown for Wasserstein GAN problems, given the defined Sobolev norm and a small enough λ the proximal objective $V_\lambda^{\text{prox}}(G, D)$ will possess a Nash equilibrium solution. This result motivates performing the minimax optimization for the proximal objective $V_\lambda^{\text{prox}}(G, D)$ instead of the original objective $V(G, D)$. Therefore, we propose proximal training in which we solve the following minimax optimization problem:

$$\min_{G_{\boldsymbol{\theta}} \in \mathcal{G}} \max_{D_{\mathbf{w}} \in \mathcal{D}} V_\lambda^{\text{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}}), \tag{18}$$

with the proximal operator defined according to the Sobolev norm in (15).

In order to take the gradient of $V_\lambda^{\text{prox}}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}})$ with respect to $\boldsymbol{\theta}$, Proposition 4 suggests solving the proximal optimization followed by computing the gradient of the original objective $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}^*})$ where the discriminator is parameterized with the optimal $\mathbf{w}^*$ to the proximal optimization.

Algorithm 1 summarizes the main steps of proximal training. At every iteration, the discriminator is optimized for T gradient steps with an additive Sobolev norm penalty forcing the discriminator to remain in the proximity of the current discriminator. Next, the generator is optimized using a gradient descent method with the gradient evaluated at the optimal discriminator solving the proximal optimization. The stepsize parameter γ_k can be adaptively selected at every iteration k. In practice, we can solve the proximal maximization problem via a first-order optimization method for a certain number of iterations. Assuming the conditions of Proposition 4 hold, the proximal optimization leads to the maximization of a strongly-concave objective which can be solved linearly fast through first-order optimization methods.

6. Numerical Experiments

To experiment the theoretical results of this work, we performed several experiments using the (Gulrajani et al., 2017)'s implementation of Wasserstein GANs with the code available at the paper's Github repository. In addition, we used the implementations of (Miyato et al., 2018; Farnia et al., 2019) for applying spectral regularization to the discriminator network. In the experiments, we used the DCGAN 4-layer CNN architecture for both the discriminator and generator functions (Radford et al., 2015) and ran each experiment for 200,000 generator iterations with 5 discriminator updates per generator update. We used the RMSprop optimzier (Hinton et al., 2012) for WGAN experiments with weight clipping or spectral normalization and the Adam optimizer (Kingma & Ba, 2014) for the other experiments.

6.1. Proximal equilibrium in Wasserstein and Lipschitz GANs

We examined whether the solutions found by Wasserstein and Lipschitz vanilla GANs represent proximal equilibria. Toward this goal, we performed similar experiments to Section 2's experiments for the WGAN-WC (Arjovsky et al., 2017), WGAN-GP (Gulrajani et al., 2017), and SN-GAN (Miyato et al., 2018) problems over the MNIST and CelebA datasets. In Section 2, we observed that after fixing the trained discriminator $D_{\mathbf{w}_{\text{final}}}$ the GAN's minimax objective $V(G_{\boldsymbol{\theta}}, D_{\mathbf{w}_{\text{final}}})$ kept decreasing when we optimized only the generator $G_{\boldsymbol{\theta}}$. In the new experiments, we similarly fixed the trained discriminator $D_{\mathbf{w}_{\text{final}}}$ resulted from the 200,000 training iterations, but instead of optimizing the GAN minimax objective we optimized the *proximal* objective defined by the norm (15) with $\lambda = 0.1$. Thus, we solved the following optimization problem initialized at $\boldsymbol{\theta}_{\text{final}}$ which denotes the parameters of the trained generator:

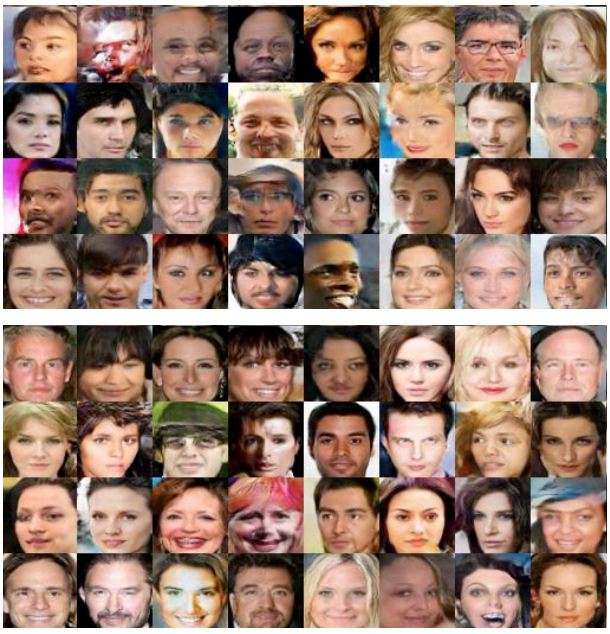

Figure 6. CelebA samples generated by SN-GAN trained via (top) regular training, (bottom) proximal training.

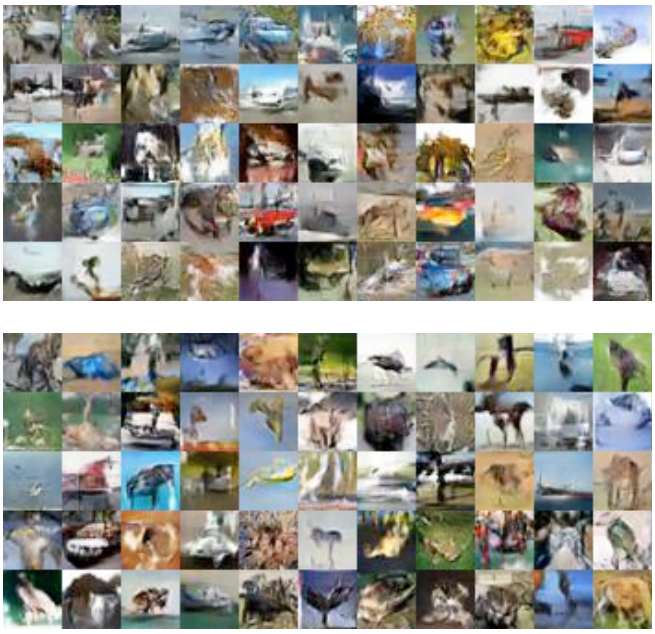

Figure 7. CIFAR-10 samples generated by SN-GAN trained via (top) regular training, (bottom) proximal training.

$$\min_{\boldsymbol{\theta}} V^{\text{prox}}_{\lambda=0.1}(G_{\boldsymbol{\theta}}, D_{\mathbf{w}_{\text{final}}}). \tag{19}$$

We computed the gradient of the above proximal objective by applying the Adam optimizer for 50 steps to approximate the solution to the proximal optimization (11) which at every iteration was initialized at $\mathbf{w}_{\text{final}}$. Figures 4 and 5 show that in the SN-GAN experiments the original minimax objective had only minor changes, compared to the results in Section 2, and the quality of generated samples did not change significantly during the optimization. We defer the similar numerical results of the WGAN-WC and WGAN-GP experiments to the Appendix. These numerical results suggest that while Wasserstein and Lipschitz GANs may not converge to local Nash equilibrium solutions as shown in Section 2, their found solutions can still represent a local proximal equilibrium.

6.2. Proximal Training improves Wasserstein and Lipschitz GANs

We applied the proximal training in Algorithm 1 to the WGAN-WC and SN-GAN problems. To compute the gradient of the proximal minimax objective, we solved the maximization problem in the Algorithm 1's first step in the for loop by applying 20 steps of Adam optimization initialized at the discriminator parameters at that iteration. Applying the proximal training to MNIST, CIFAR-10, and CelebA datasets, we qualitatively observed qualitatively better generated pictures. Figures 6 and 7 show the samples generated by SN-GAN trained on CelebA and CIFAR-10 data via proximal and regular training schemes, which look visually better for proximal training. Due to the limited space, we postpone the generated samples for WGAN experiments to the Appendix.

To quantitatively compare the proximal and regular GAN training methods, we measured the Inception scores (Salimans et al., 2016) of the samples generated in the CIFAR-10 experiments. As shown in Table 1, proximal training results in improved inception scores. In this table, DIM stands for the dimension parameter of the DC-GAN's convolutional networks.

Table 1. Inception scores for regular vs. proximal training

GAN EXPERIMENT	REGULAR	PROXIMAL
WGAN-WC (DIM=64)	4.16 ± 0.15	4.56 ± 0.19
WGAN-WC (DIM=128)	2.52 ± 0.12	4.23 ± 0.15
SN-GAN (DIM=64)	5.12 ± 0.25	5.72 ± 0.22
SN-GAN (DIM=128)	5.62 ± 0.23	6.12 ± 0.22

Acknowledgements

This work is supported by the MIT-Air Force AI Accelerator (AIAA) under grant FA8750-19-2-1000 and MIT-IBM Watson AI Lab. The authors would also like to thank the anonymous reviewers for their constructive feedback.

References

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 224–232, 2017.

Bailey, J. P. and Piliouras, G. Multiplicative weights update in zero-sum games. In *Proceedings of the 2018 ACM Conference on Economics and Computation*, pp. 321–338, 2018.

Berard, H., Gidel, G., Almahairi, A., Vincent, P., and Lacoste-Julien, S. A closer look at the optimization landscapes of generative adversarial networks. *arXiv preprint arXiv:1906.04848*, 2019.

Berthelot, D., Schumm, T., and Metz, L. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

Brock, A., Donahue, J., and Simonyan, K. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Cheung, Y. K. and Piliouras, G. Vortices instead of equilibria in minmax optimization: Chaos and butterfly effects of online learning in zero-sum games. *arXiv preprint arXiv:1905.08396*, 2019.

Daskalakis, C. and Panageas, I. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pp. 9236–9246, 2018.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Farnia, F. and Tse, D. A convex duality framework for gans. In *Advances in Neural Information Processing Systems*, pp. 5248–5258, 2018.

Farnia, F., Zhang, J., and Tse, D. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2019.

Fedus, W., Rosca, M., Lakshminarayanan, B., Dai, A. M., Mohamed, S., and Goodfellow, I. Many paths to equilibrium: Gans do not need to decrease a divergence at every step. *arXiv preprint arXiv:1710.08446*, 2017.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. Understanding gans: the lqg setting. *arXiv preprint arXiv:1710.10793*, 2017.

Fiez, T., Chasnov, B., and Ratliff, L. J. Convergence of learning dynamics in stackelberg games. *arXiv preprint arXiv:1906.01217*, 2019.

Flokas, L., Vlatakis-Gkaragkounis, E.-V., and Piliouras, G. Poincar\'e recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. *arXiv preprint arXiv:1910.13010*, 2019.

Goodfellow, I. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

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*, pp. 2672–2680, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 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*, pp. 6626–6637, 2017.

Hinton, G., Srivastava, N., and Swersky, K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. 14(8), 2012.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002*, 2018.

Jin, C., Netrapalli, P., and Jordan, M. I. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kodali, N., Abernethy, J., Hays, J., and Kira, Z. On convergence and stability of gans. *arXiv preprint arXiv:1705.07215*, 2017.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Lei, Q., Lee, J. D., Dimakis, A. G., and Daskalakis, C. Sgd learns one-layer networks in wgans. *arXiv preprint arXiv:1910.07030*, 2019.

Lin, T., Jin, C., and Jordan, M. I. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

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

Mazumdar, E. V., Jordan, M. I., and Sastry, S. S. On finding local nash equilibria (and only local nash equilibria) in zero-sum games. *arXiv preprint arXiv:1901.00838*, 2019.

Mertikopoulos, P., Papadimitriou, C., and Piliouras, G. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 2703–2717. SIAM, 2018.

Mescheder, L., Nowozin, S., and Geiger, A. The numerics of gans. In *Advances in Neural Information Processing Systems*, pp. 1825–1835, 2017.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 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. *arXiv preprint arXiv:1901.08511*, 2019.

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., and Cheng, Y. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

Nagarajan, V. and Kolter, J. Z. Gradient descent gan optimization is locally stable. In *Advances in neural information processing systems*, pp. 5585–5595, 2017.

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*, pp. 14905–14916, 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*, pp. 271–279, 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Roth, K., Lucchi, A., Nowozin, S., and Hofmann, T. Stabilizing training of generative adversarial networks through regularization. In *Advances in neural information processing systems*, pp. 2018–2028, 2017.

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*, pp. 2234–2242, 2016.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pp. 7091–7101, 2018.

Schäfer, F., Zheng, H., and Anandkumar, A. Implicit competitive regularization in gans. *arXiv preprint arXiv:1910.05852*, 2019.

Taghvaei, A. and Jalali, A. 2-wasserstein approximation via restricted convex potentials with application to improved training for gans. *arXiv preprint arXiv:1902.07197*, 2019.

Thekumparampil, K. K., Jain, P., Netrapalli, P., and Oh, S. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pp. 12659–12670, 2019.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In *International Conference on Machine Learning*, pp. 6586–6595, 2019.

Wang, Y., Zhang, G., and Ba, J. On solving minimax optimization locally: A follow-the-ridge approach. In *International Conference on Learning Representations*, 2020.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

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*, pp. 11598–11610, 2019.

Zhou, Z., Liang, J., Song, Y., Yu, L., Wang, H., Zhang, W., Yu, Y., and Zhang, Z. Lipschitz generative adversarial nets. *arXiv preprint arXiv:1902.05687*, 2019.