Implicit competitive regularization in GANs

Florian Schäfer
Caltech
schaefer@caltech.edu

Hongkai Zheng
Shanghai Jiao Tong University *
devzhk@sjtu.edu.cn

Anima Anandkumar
Caltech
anima@caltech.edu

Abstract

While generative adversarial networks (GANs) are capable of producing high quality samples, they suffer from instability and mode collapse during training. To alleviate this problem, multiple authors have proposed gradient penalties on the discriminator, which are thought of as replacing the minimization of the Jensen-Shannon (JS) divergence implicit in the original GAN framework with minimization of the Wasserstein distance. In the present work, we provide a mechanism, *implicit competitive regularization*, by which rational play of the two players' can stabilize the dynamics of a GAN, even if the Jensen-Shannon divergence is not meaningful. We furthermore observe that competitive gradient descent (CGD) (Schäfer and Anandkumar, 2019) can take advantage of this mechanism in order to achieve stable GAN training, without imposing gradient penalties. Our numerical experiments suggest that GAN training with CGD is stable and only needs regularization to prevent overfitting, to a similar extent as ordinary neural networks.

1 Introduction

Generative adversarial networks (GANs): (Goodfellow et al., 2014) are powerful generative models that constitute the state of the art on a variety of problems. Whereas other methods, like variational autoencoders (VAEs) (Kingma and Welling, 2013; Rezende et al., 2014), optimize target functionals given a priori, GANs are constructed from a competitive game between a generator that tries to generate realistic samples, and a discriminator that tries to distinguish real data from artificial samples. Both players are represented as neural networks, each of which iteratively updates its parameters to improve their respective loss functions. While GANs are capable of producing high quality samples, their training is known to suffer from instability and *mode collapse*, a phenomenon whereby the diversity of the produced samples drops dramatically.

Unbiased loss functions: In order to define a game between generator and discriminator, we need to define their respective loss functions. The original GAN paper suggested to use the relative cross entropy loss function, resulting in the zero-sum or minimax game

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}}\left[\log \mathcal{D}(x)\right] + \frac{1}{2}\mathbb{E}_{x\sim \mathcal{G}}\left[\log\left(1-\mathcal{D}(x)\right)\right]. \tag{1}$$

This is the first example of what we call an *unbiased loss function*, which more generally leads to games of the form

$$\min_{\mathcal{G}} \mathbb{E}_{x\sim P_{\text{data}}}\left[f_{\text{real}}\left(\mathcal{D}\left(x\right)\right)\right] + \mathbb{E}_{x\sim \mathcal{G}}\left[f_{\text{fake}}\left(\mathcal{D}\left(x\right)\right)\right] \tag{2}$$

$$\min_{\mathcal{D}} \mathbb{E}_{x\sim P_{\text{data}}}\left[g_{\text{real}}\left(\mathcal{D}\left(x\right)\right)\right] + \mathbb{E}_{x\sim \mathcal{G}}\left[g_{\text{fake}}\left(\mathcal{D}\left(x\right)\right)\right]. \tag{3}$$

*This work was produced while HZ was a visiting undergraduate researcher at Caltech.

Abstract for: Smooth Games Optimization and Machine Learning Workshop (NeurIPS 2019), Vancouver, Canada.

We call these loss functions unbiased since they only evaluate the generator and discriminator in terms of their performance in generating and identifying synthetic samples. Other examples of this class are the nonsaturating loss (Goodfellow et al., 2016)[Section 3.2.3], f-GAN (Nowozin et al., 2016), and LSGAN (Mao et al., 2017).

Biased loss functions: When taking the maximum over all possible $\mathcal{D}$, Equation (1) amounts to minimization of the Jensen-Shannon (JS) divergence between $\mathcal{G}$ and P_{data}. This has motivated the interpretation of the original GAN loss as an approximate minimization of the JS divergence where we approximate the inner optimization over $\mathcal{D}$ by making a finite number of gradient ascent updates on $\mathcal{D}$. Beginning with Arjovsky and Bottou (2017), a number of works have used this point of view to explain the instability and mode collapse in GAN training (Arjovsky et al., 2017; Roth et al., 2017; Arora et al., 2017). Since P_{data} is generally atomic (finite data) and often concentrated on low-dimensional structures while $\mathcal{G}$ is the differentiable push-forward of a continuous distribution, P_{data} and $\mathcal{G}$ will usually be mutually singular, leading to a maximal JS divergence that does not yield useful gradient information. Indeed, this pathology is present for all unbiased loss functions since if the support of P_{data} has measure zero under $\mathcal{G}$, a sufficiently powerful discriminator can fully minimize $\mathbb{E}_{x\sim P_{\text{data}}}\left[g_{\text{real}}\left(\mathcal{D}\left(x\right)\right)\right]$ without compromising the minimization of $\mathbb{E}_{x\sim\mathcal{G}}\left[g_{\text{fake}}\left(\mathcal{D}\left(x\right)\right)\right]$. To resolve this issue, Arjovsky and Bottou (2017); Arjovsky et al. (2017) introduce Wasserstein GAN (WGAN) as the solution to the minimax problem

$$\min_{\mathcal{G}}\max_{\|\mathcal{D}\|_1\leq 1}\mathbb{E}_{x\sim P_{\text{data}}}\left[\mathcal{D}(x)\right]-\mathbb{E}_{x\sim P_{\mathcal{G}}}\left[\mathcal{D}(x)\right],$$

where $\|\cdot\|_1$ denotes the Lipschitz seminorm. For a perfect discriminator, the above quantity is equal to the Wasserstein distance, which provides a nontrivial notion of distance even among mutually singular measures. While the initial WGAN imposes the Lipschitz constraint by using weight clipping, Gulrajani et al. (2017) proposed WGAN gradient penalty (WGAN-GP) that approximately enforces it by penalizing the gradient of $\|\nabla\mathcal{D}\|$. Subsequently, a number of similar penalties have been introduced Arora et al. (2017); Kodali et al. (2017); Miyato et al. (2018); Adler and Lunz (2018); Mroueh et al. (2017). The games resulting from these losses have the general form

$$\min_{\mathcal{G}}\mathbb{E}_{x\sim P_{\text{data}}}\left[f_{\text{real}}\left(\mathcal{D}\left(x\right)\right)\right]+\mathbb{E}_{x\sim\mathcal{G}}\left[f_{\text{fake}}\left(\mathcal{D}\left(x\right)\right)\right]+F(\mathcal{G},\mathcal{D}) \tag{4}$$

$$\min_{\mathcal{D}}\mathbb{E}_{x\sim P_{\text{data}}}\left[g_{\text{real}}\left(\mathcal{D}\left(x\right)\right)\right]+\mathbb{E}_{x\sim\mathcal{G}}\left[g_{\text{fake}}\left(\mathcal{D}\left(x\right)\right)\right]+G(\mathcal{G},\mathcal{D}), \tag{5}$$

where $F(\mathcal{G},\mathcal{D}), G(\mathcal{G},\mathcal{D})$ might take on the value ∞ if a constraint, like $\|\mathcal{D}\|_1\leq 1$, is violated. We call these loss functions biased loss functions, since the terms $F(\mathcal{G},\mathcal{D})$ and $G(\mathcal{G},\mathcal{D})$ express a preference for certain $\mathcal{G}$ and $\mathcal{D}$ that need not depend on their performance in generating and identifying synthetic samples.

Is bias necessary? Biased loss functions based on gradient penalties have become popular among practitioners as ways of mitigating instability and mode collapse. However, they are dissatisfying from a conceptual perspective since they require the choice of a metric on sample space to measure the size of gradients. In practice, this metric is typically chosen as ℓ_2, which is clearly a poor notion of similarity if the datapoints are for example images of human faces. Unbiased loss functions, in contrast, have the intriguing property that they do *not* require a choice of metric! Rather, the notion of similarity underlying the samples of GANs with unbiased loss functions arises from the subtle interplay of architectural "implicit biases" and the use of stochastic gradient based optimization. Based on the observations that neural networks seem to be reasonably good at imitating human perception, we believe that it is this feature that allows GANs to produce much sharper and realistic images than any other method. Therefore, it is desirable to mitigate the instability of unbiased loss functions without artificially imposing a metric on sample space.

Our contribution: In this work we investigate whether the stabilization of GANs requires augmenting the loss function with additional penalties or whether stable algorithms can be obtained even for unbiased loss functions. By studying a toy example we observe that if the two players have limited information of the optimization landscape and are aware of the competitive nature of the game, rational play can lead to stable behavior even if either player can receive arbitrarily large rewards by diverging towards infinity. We then argue empirically that key features of our toy problem are present in real GANs, and rational play can therefore mitigate the pathological behavior observed by Arjovsky and Bottou (2017) without adding additional regularization. We then point out that the updates of competitive gradient descent (CGD) introduced by Schäfer and Anandkumar (2019) incorporate a suitable notion of rational play as outlined in the toy example. Under competitive gradient descent,

the two agents prefer strategies that are robust to the actions of each other, which greatly increases the stability of the resulting algorithm. This leads to a form of regularization that does not require the choice of a metric on sample space and which we call *implicit competitive regularization*.

2 Implicit competitive regularization

What is the solution of a GAN?: The main objection to the original GAN loss has been that even for a well-trained generator there exist discriminators that can distinguish it arbitrarily well from the true data, making the minimax interpretation of GANs meaningless. While Arora et al. (2017) pointed out that the set of relevant discriminators is limited by the capacity of the discriminator we observe experimentally, just as Arjovsky and Bottou (2017), that even capable generators can incur large losses under a discriminator that is optimized with respect to this particular generator. While Kodali et al. (2017), as well as multiple works on the algorithmic aspects of GAN training Daskalakis et al. (2017); Daskalakis and Panageas (2018); Jin et al. (2019) emphasize the view of GANs as *local* Nash equilibria, our experiments in Figure 2 suggest the discriminator can still be improved significantly with a small number of updates. Berard et al. (2019) further observed that in many good GAN solutions the generator is not at a local minimum, casting further doubt at the interpretation of GANs as seeking local Nash equilibria. We will now illustrate a mechanism by which the rational actions of players under limited information can even stabilize pairs of globally *worst* strategies.

A simple toy problem: Consider the zero-sum game given by

$$\min_{x\in\mathbb{R}} \max_{y\in\mathbb{R}} -\exp(x^2) - \alpha xy + \exp(y^2) \tag{6}$$

where $\alpha \gg 1$ is a large but fixed parameter. This game does not have a global Nash equilibrium, since each player can always achieve an exponentially increasing reward by moving towards infinity. It furthermore does not have local Nash equilibria, since the curvature of the objective function is always negative in x and positive in y.

Despite the absence of equilibria in the classical sense, we could hope that a suitable iterative play by the two players has stable behavior. Let us assume to this end that x starts in 0 and y in 2 and that at each round, both players are allowed to change their strategy by at most distance one. If the players aim to mimimize their cumulative loss over the course of the game, a winning strategy of y is still to move towards $+\infty$ as quickly as possible (see Figure 1). Since limiting the size of the players steps did not suffice to stabilize their dynamics, we could now limit the players to only using local information. We assume that at the $k-th$ step, the player x (y) only has access to to objective funcition on $[x_k-1, x_k+1]\times\{y\}$ ($\{x\}\times\{y_k-1, y_k+1\}$) and tries to minimize their loss in the next round assuming, for the lack of additional information, that the loss function will stay the same. As shown in Figure 1, the strategies initially exhibit oscillations that slow down divergence until evenually the exponential terms dominate (see Figure 1). This behavior is similar to the oscillations of simultaneous gradient descent and arises from both players being ignorant of each other's presence. Alternatively, we can consider the situation where at each step, both players have access to the loss function on $[x_k-1, x_k+1]\times[y_k-1, y_k+1]$ and try to minimize their loss, aware that the other player tries to minimize theirs. In our toy example, the resulting optimal strategies are deterministic and result in stable dynamics (see Figure 1).

Optimal strategies in GANs are brittle: Even though both players were able to achieve arbitrarily low losses by moving off to infinity in the above toy example, the combination of iterative play, limited information, and strategic behavior lead to stable dynamics. We will now investigate whether a similar mechanism could be at play in allowing GANs to achieve meaningful solutions even though for any given generator $\mathcal{G}$ there exists a discriminator $\mathcal{D}$ that achieves a loss of maximal JS divergence. The key feature at play in the toy example was that as any player moves towards infinity, it the resulting strategy becomes more and more vulnerable to counterplay of the other player. The corresponding behavior in GANs would be that as $\mathcal{D}$ achieves better and better performance against a given discriminator, it also becomes more and more vulnerable to changes in the generator's strategy. Indeed, Figure 2 shows that the near-optimal discriminators of the experiment in Figure 2 incur large losses, after just a few iterations of training the generator. Thus, even without regularization, a rational discriminator might refrain from fully maximizing the loss. However, as we have seen in the toy example, naive gradient dynamics can nevertheless lead to unstable behavior and divergence, leading us to suspect that the reason for instability in GANs is not an inadequate loss function but rather the overly naive notion of strategic play modelled by simultaneous or alternating gradient descent.

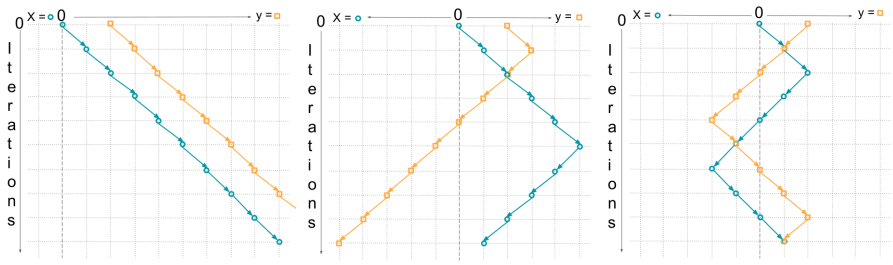

Figure 1: **Competitive regularization in toy problem:** Under full information, each player moves towards infinity as quickly as possible (first panel). Under limited information, but without accounting each other's actions, the players oscillate and eventually diverge (second panel). Under both limited information and awareness of the opponent, the trajectories become stable (third panel).

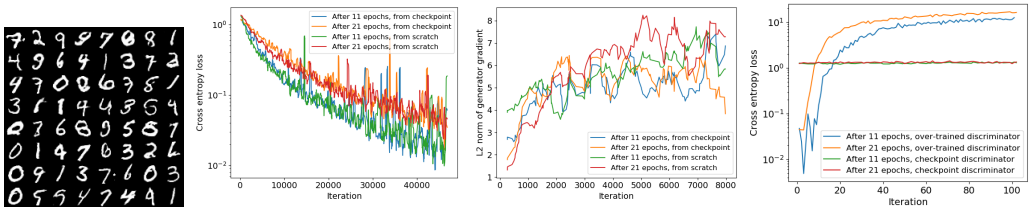

Figure 2: **Overtraining Discriminators:**We begin by training a GAN on MNIST using Adam for 21 (11) epochs, saving the resulting generator (samples in first panel) and discriminator as a "checkpoint". We then *overtrain* the discriminator while keeping the generator of the checkpoint fixed, achieving low discriminator loss (second panel). As indicated by the increasing gradient norm of the generator during training (third panel), the resulting discriminator is brittle and incurs large losses as we start training the generator again (fourth panel).

Competitive gradient descent uses implicit competitive regularization: As opposed to the toy example, in gradient based optimization the player don't have access to the full objective function within a neighborhood, but rather to a local taylor approximation of the loss function. As this Taylor approximation gradually becomes unreliable as we move away from where we computed the derivatives, we add an additional regularization term when determining the update step *Competitive gradient descent* (Schäfer and Anandkumar, 2019) to use the Nash equilibrium of a quadratically regularized local bilinear approximation to the game as updates at each step. In the special case of a zero-sum game with loss of f for player x, this amounts to the update rule

$$x_{k+1} = x_k - \eta \left(\mathrm{Id} + \eta^2 D^2_{xy} f D^2_{yx} f\right)^{-1} \left(\nabla_x f + \eta D^2_{xy} f \nabla_y f\right)$$
$$y_{k+1} = y_k + \eta \left(\mathrm{Id} + \eta^2 D^2_{yx} f D^2_{xy} f\right)^{-1} \left(\nabla_y f - \eta D^2_{yx} f \nabla_x f\right).$$

Here, $\nabla_x f$ corresponds to gradient descent in x, while $\eta D^2_{xy} f \nabla_y f$ models the anticipation of y playing gradient descent. The term $\left(\mathrm{Id} + \eta^2 D^2_{xy} f D^2_{yx} f\right)^{-1}$ corresponds to avoiding brittle updates, the result of which strongly depends on the the actions of the other players. In the limit of large $D_{xy} f$ this restricts updates to the subspace orthogonal to the singular vectors of $D_{xy} f$. We observe in practice that this additional regularization, which we refer to as *implicit competitive regulariztation* (ICR), can stabilize the dynamics of GAN training. This is remarkable, since ICR does not require the introduction of additional biases but emerges from the subtle interplay of architecture and strategic behavior of the two networks.

Conclusion: Our empirical results suggest that training GANs with CGD can lead to stable training dynamics even in the absence of any form of regularization. While adding some regularization like dropout, early stopping, noise, weight decay, or gradient penalty might be necessary to prevent overfitting (just like in ordinary neural networks), we argue that when using CGD there is no fundamental lack of stability that needs to be addressed by a particular modification of the loss function.

Acknowledgments

A. Anandkumar is supported in part by Bren endowed chair, DARPA PAIHR00111890035, Raytheon, and Microsoft, Google and Adobe faculty fellowships. F. Schäfer gratefully acknowledges support by the Air Force Office of Scientific Research under award number FA9550-18-1-0271 (Games for Computation and Learning). H. Zheng is supported by Zhiyuan College, Shanghai Jiao Tong University.

References

Adler, J. and Lunz, S. (2018). Banach wasserstein gan. In *Advances in Neural Information Processing Systems*, pages 6754–6763.

Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223.

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

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

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

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

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT Press.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

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

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

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*.

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

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*, pages 271–279.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

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

Schäfer, F. and Anandkumar, A. (2019). Competitive gradient descent. *arXiv preprint arXiv:1905.12103*.