Understanding Generative Adversarial Networks from First Principles

Weili Nie, Ankit Patel
Electrical and Computer Engineering, Rice University

Abstract

Generative adversarial networks (GANs) are powerful generative models and have been very successful in many deep learning tasks, but they suffer from training issues such as vanishing or exploding gradients and mode collapse. To analyze these issues, we first build a theoretical framework to derive the original form of vanilla GAN by applying six key assumptions. Based on this framework, we discuss the possible sources of training issues in GANs with relaxations of those assumptions. We show that for general case, it is better to analyze GANs from game theoretical perspective rather than minimizing any divergence or distance. Furthermore, we prove the existence of Nash equilibrium in the ideal vanilla GAN and some of its variants, and then analyze a specific case where nonconvergecen may exist even if there is a Nash equilibrium. Experimental results on both synthetic dataset and celebA dataset verify our theoretical analysis.

1 Motivations

Generative Adversarial Networks (GANs) [1] have achieved great success at generating realistic and sharp looking images, with their applications to many important problems, such as semi-supervised learning, reinforcement learning, text-to-image translation, and creating high resolution images, etc. However, they still remain highly difficult to train. Most approaches to deal with this problem rely on heuristics with little theory explaining the training instability or non-convergence behavior of GANs. Furthermore, the reason that GANs produce realistically looking samples is not entirely clear. The lack of theoretical understanding of GANs has prohibited their safer and more reliable uses in dangerous applications such as self-driving cars.

Therefore, the objective of our work is to provide a solid theoretical understanding of these issues, and to create principled methods towards addressing them. In particular, we are interested in looking into our experimental results while training GANs, and understanding what assumptions have gone into GANs while trying to explain theoretically our observations.

2 Setup and Key Assumptions in GANs

First of all, we would like to build a mathematical framework that systematically proposes how to get the original form of vanilla GAN [1]. Based on some key assumptions, we derive its objectives (1) as shown below from the divergence minimization and likelihood ratio estimation perspective, which gives more flexibilities in interpreting GANs objectives.

$$\min_{\phi} \max_{\theta} \mathbb{E}_{x\sim\mathbb{P}_r} \left[\log D_\theta\left(x\right)\right] + \mathbb{E}_{z\sim\mathbb{P}_z} \left[\log\left(1 - D_\theta\left(G_\phi(z)\right)\right)\right] \tag{1}$$

where $\mathbb{P}_r$ and $\mathbb{P}_z$ are the real data distribution and noise distribution, respectively. $D_\theta(x)$ is the discriminator with parameters θ that represents the probability that x comes from the real data distribution, and $G_\phi(z)$ is the generator with parameters ϕ that maps the noise to data space. Also,

$G_\phi(z) \sim \mathbb{P}_g$ which represents generation distribution. The ultimate goal of the objective (1) is to learn parameters θ and ϕ such that $\mathbb{P}_g = \mathbb{P}_r$.

Assumption 1. *We assume that the real data x and generated data $G_\phi(z)$ are absolutely continuous random variables, and thus vanilla GAN starts from minimizing f-divergence [2].*

From the Radon-Nikodym theorem, under this assumption both real data x and generated data $G_\phi(z)$ have well-defined probability density functions, which are denoted by $p_r(x)$ and $p_g(x)$, respectively. Then the f-divergence between $\mathbb{P}_g$ and $\mathbb{P}_r$ is defined as $D_f\left(\mathbb{P}_r \| \mathbb{P}_g\right) = \mathbb{E}_{x\sim\mathbb{P}_g} f\left(\frac{p_r(x)}{p_g(x)}\right)$ where f is a convex function satisfying $f(1) = 0$.

Assumption 2. *Given generator G_ϕ, discriminator D_θ is a binary classifier using cross-entropy objective to estimate the likelihood ratio as $\Lambda(x) = \frac{p_r(x)}{p_g(x)}$.*

According to Bayes rule, we have $P\left(c = 1 \,|x\right) = \frac{\Lambda(x)}{\Lambda(x)+1}$, so likelihood ratio estimation becomes a binary classification problem.

Assumption 3. *Denote by $D_\theta\left(x\right)$ a parametrized neural network which maps data from $\mathcal{X}$-space to $[0, 1]$. Assume it has enough expressive power to approximate the posterior class probability $P\left(c = 1 \,|x\right)$.*

Based on this assumption, we can replace $P\left(c = 1 \,|x\right)$ by $D_\theta\left(x\right)$ and thus minimizing the cross-entropy with respect to parameters θ is equivalent to maximizing the discriminator objective in the vanilla GAN.

Assumption 4. *Denote $G_\phi(z)$ by a parametrized neural network which maps noise with known density $z \sim \mathbb{P}_z$ (e.g. Gaussian distribution) from $\mathcal{Z}$-space to $\mathcal{X}$-space. Assume it has enough expressive power to approximate the real data distribution $\mathbb{P}_r$.*

Assumption 5. *There are many forms of f-divergence, and the generator objective (1) comes from setting f to be the form associated to JS divergence.*

After some algebraic manipulations, the above two assumptions result in the minimax objective (1) in the vanilla GAN. In reality, however, we use only finite samples followed from $\mathbb{P}_r$ and $\mathbb{P}_z$ to estimate the expectations. Thus we are actually optimizing the empirical objective, which relies on an additional assumption.

Assumption 6. *We have sufficiently large training data to accurately estimate the expectations in the minimax objective (1).*

In the following, we will see how strong the above assumptions actually are and/or how relaxing them impacts the training instability and non-convergence, mode collapse, generalization, etc.

3 Analysis of Key Issues in GANs

3.1 Different Theoretical Objectives

For KL-divergence, consider an extreme case that the two distributions only differ in the neighborhood $B_\epsilon(x_0)$ of a point x_0. If $p_r(x) > 0$ and $p_g(x) \to 0$ for $x \in B_\epsilon(x_0)$, the KL will go to infinity. This means that KL divergence assigns *a high cost for mode collapse* where $\mathbb{P}_g$ does not cover the whole support of $\mathbb{P}_r$. On the other hand, if $p_r(x) \to 0$ and $p_g(x) > 0$ for $x \in B_\epsilon(x_0)$, the KL will go to zero. This means that KL divergence *assigns a low cost for generating implausible samples*. Just the opposite, reverse KL assigns a low cost for mode collapse and a high cost for generating implausible samples. So reverse KL is a desirable objective to generate realistic samples but it will suffer from mode collapse.

Unlike KL and reverse KL, JS divergence which is implicitly in vanilla GAN is symmetrical and thus it will equally favor of both generating realistic samples and covering the data support.

3.2 Samples in Lower Dimensional Manifolds

We can relax Assumption 1 by considering the realistic situation that data distribution $\mathbb{P}_r$ and generation distribution $\mathbb{P}_g$ are both supported by lower dimensional manifolds. Particularly, Arjovsky

et at. [3] proved that if the dimension of input noise space is lower than that of data space, which is actually the case during most GANs training, $\mathbb{P}_g$ generated by a neural network will be supported by a lower dimensional manifold in the data space.

Equivalently, it means that density functions of $\mathbb{P}_r$ and $\mathbb{P}_g$ are not well defined, and the divergences we used in last subsection will be either constant or infinite, but not continuous with respect to parameters ϕ when the two manifolds do not align perfectly. So after the discriminator is trained to the optimal, gradient descent via backpropogation will not work any more. Intuitively, JS divergence which is a constant corresponds to vanishing gradient problems in vanilla GAN, and reverse KL divergence which is infinite relates to training instability problems.

3.3 Finite Samples and Generalization

If we relax Assumption 6 by considering finite training samples, the empirical objective might not be a good approximation of the original form (1). That is, the *generalization* of GANs training might not be guaranteed. In particular, Arora et al. [4] shows that even if the generator happens to find the real distribution, if the discriminator has sufficient expressive power to compute either JS divergence or Wasserstein distance [5], then the distance between empirical distributions is still large and the generator has no idea that it has succeeded and might move away [4]. In this sense, if the number of training samples is not sufficiently large, JS divergence and even Wasserstein distance are no longer desirable metrics as GANs objective.

3.4 Expressive Power and Convergence

we can further relax Assumption 3 and 4 by considering limited expressive power of discriminator and generator. In this case, it is more suitable to interpret GANs training as a two-player game than minimizing an approximate divergence or distance.

First, suppose generator has finite capacity, but discriminator is perfect. Normally, discriminator will win the game, and the generated sample quality depends on the expressive power of generator. The only chance the generator can win might be by memorizing the real training samples, which is the result of overfitting. Second, suppose generator is perfect, but discriminator has finite capacity. Generator will win the game, but with high probability, generator cannot find the exact target data manifold since the gradient vanishes before reaching the optimum. Third, suppose both generator and discriminator have finite capacity, which is more realistic during GANs training. In this case, it is unclear whether the two-player game could converge to any equilibrium via backpropagation and whether generator could win the game.

4 Conclusions

We first built a theoretical framework to derive the original form of vanilla GAN by applying some key assumptions. Based on this framework, we discussed the issues in GANs such as vanishing gradient, unstable training and mode collapse with relaxations of those assumptions.

References

[1] I. Goodfellow, *et al.* "Generative Adversarial Nets." in *NIPS*, 2014.

[2] S. Nowozin, B. Cseke, and R. Tomioka, "f-GAN: Training generative neural samplers using variational divergence minimization." In *NIPS*, 2016.

[3] M. Arjovsky and L. Bottou. "Towards principled methods for training generative adversarial networks." in *ICLR*, 2017.

[4] S. Arora, et al. "Generalization and equilibrium in generative adversarial nets (GANs)." *arXiv preprint arXiv:1703.00573*.

[5] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN." *arXiv preprint arXiv:1701.07875 (2017)*.