Convergence Problems with Generative Adversarial Networks (GANs)

S. A. Barnett
MMathPhil Mathematics and Philosophy

A dissertation presented for
CCD Dissertations on a Mathematical Topic

Mathematical Institute
University of Oxford
Hilary Term 2018

Abstract

Generative adversarial networks (GANs) are a novel approach to generative modelling, a task whose goal it is to learn a distribution of real data points. They have often proved difficult to train: GANs are unlike many techniques in machine learning, in that they are best described as a two-player game between a discriminator and generator. This has yielded both unreliability in the training process, and a general lack of understanding as to *how* GANs converge, and if so, to *what*. The purpose of this dissertation is to provide an account of the theory of GANs suitable for the mathematician, highlighting both positive and negative results. This involves identifying the problems when training GANs, and how topological and game-theoretic perspectives of GANs have contributed to our understanding and improved our techniques in recent years.

Acknowledgements

This work was originally presented for the Part C Dissertations on a Mathematical Topic in the MMathPhil Mathematics and Philosophy course at the University of Oxford. As such, I would like to first and foremost thank my supervisor Varun Kanade for his constructive feedback and support in writing this. I would also like to thank my tutor at Worcester College, Richard Earl, for the support during this project and throughout my time at Oxford.

Contents

Chapter 1

Introduction

Generative adversarial networks (GANs) were proposed by Goodfellow et al. (2014) as a novel approach to generative modelling, a task whose goal it is to learn a distribution of real data points.

The term *adversarial* refers to the use of two opposing neural networks in GANs: a *discriminator* trained to tell real data samples apart from GAN-produced samples, and a *generator* that seeks to fool the discriminator. As can be seen in Figure 1.1, GANs are capable of producing stunningly realistic samples.

However, they have also proved difficult to train: GANs are unlike many techniques in machine learning, in that they are best described as a two-player game between a discriminator and generator. This has yielded both unreliability in the training process, and a general lack of understanding as to *how* GANs converge, and if so, to *what*.

The purpose of this dissertation is to provide an account of the theory of GANs suitable for the mathematician, highlighting both positive and negative results. Chapter 2 introduces GANs in their original formulation, in addition to some of the main problems encountered during the training process. The results of this chapter are largely due to Goodfellow et al. (2014), though I provide for the first time an explicit proof[1] of Proposition 2.6.1, and give a novel example (Corollary 2.6.2) of its negative consequences for the training of GANs. Chapter 3 gives a perspective of GANs as minimising some divergence between the generator distribution and the target distribution, arguing that certain variants of GANs may induce a more practically useful notion of divergence than that induced by the original GAN formulation. The results of this chapter are predominately due to Arjovsky et al. (2017). Chapter 4 discusses GANs from the perspective of game theory, which allows for a broader modelling of GAN training dynamics than that of the previous chapter. The change in emphasis is inspired by Fedus et al. (2017), with the main result for GANs coming from Arora et al. (2017). Chapter 5 concludes this work.

1.1 Notation

- $\mu \otimes \nu$ - the product measure, for measures μ and ν.
- $\mathcal{B}(\mathcal{X})$ - the space of Borel-measurable subsets of $\mathcal{X}$.

[1] The result has been claimed, though not proven, by a number of authors (Goodfellow et al., 2014; Goodfellow, 2016; Metz et al., 2016; Arjovsky et al., 2017).

Figure 1.1: 1024×1024 images generated on the CelebA-HQ dataset. Image taken from Karras et al. (2017, Figure 5).

- $\text{Prob}(\mathcal{X})$ - the space of probability measures on $\mathcal{X}$.
- $\mathcal{C}_b(\mathcal{X})$ - the space of bounded, continuous functions from $\mathcal{X}$ to $\mathbb{R}$.
- $[N]$ - the set of integers $\{1, ..., N\}$, where $N \in \mathbb{N}$.
- $\mathcal{N}(\mu, \sigma^2)$ - the Gaussian distribution on $\mathbb{R}^n$ with mean μ and variance σ^2.

Chapter 2

GANs: Initial Results

2.1 Motivation

The goal of generative modelling is to learn a particular distribution p_r of real data. The distribution p_r may be represented explicitly[1], or *implicitly* by providing a means to produce samples from the distribution.

The latter approach is taken by **generative adversarial networks (GANs)**. We may view a GAN as a game between two players: a **generator** and a **discriminator**. The former is represented by a function G inducing a distribution p_G (see below), and the latter by a function D.

Consider a GAN trained on images of people. Given a fixed generator, the discriminator is trained to distinguish between images produced by the real dataset (labelled 1) and images produced by the generator (labelled 0). It does so by mapping each data point x to a value in $[0, 1]$. In some sense, $D(x)$ represents the probability that x was a real sample rather than a generated one.

The goal of the generator, consequently is to produce images that the discriminator will classify as being real, while the goal of the discriminator is to classify these same images as in fact being produced by the generator. The generator induces a distribution p_G by taking a sample z from a (typically Gaussian) prior on input noise variables, p_z, and mapping it to a synthetic data point $G(z)$. Therefore, if $z \sim p_z$, then p_G is the distribution such that $G(z) \sim p_G$.

The generator aims to produce points such that $D(G(z))$ is closer to 1. In other words, the generator is trying to fool the discriminator. As we shall see, it will succeed at doing so when the images it produces arise from the same probability distribution as that of the real images of people.

2.2 Minimax-GANs

To formally specify a GAN, we need to give to the generator and the discriminator an objective that each seeks to optimise. Though we may give the discriminator and generator distinct objectives, it is common and often useful for there to be one objective that the generator seeks to **mini**mise, and that the discriminator seeks to **max**imise. In this case, which I shall refer to as the **minimax**[2] case, we can

[1] For a taxonomy of such methods, see Goodfellow (2016, Sections 2.2-2.4).

[2] This minimax perspective will be elaborated in the context of game theory in Chapter 4.

represent the objective by a single **value function** $V(D, G)$.

Definition 2.2.1 (Idealised MM-GANs (Goodfellow et al., 2014))**.** Let $\mathcal{Z} \subseteq \mathbb{R}^\ell$, $\mathcal{X} \subseteq \mathbb{R}^d$ be ambient data spaces, let p_z be a prior distribution over $\mathcal{Z}$, and let p_r be the distribution of real data points over $\mathcal{X}$. The **idealised minimax GAN** (IMM-GAN) is the game specified by the objective

$$\min_G \max_D V_{\text{IMM}}(D, G), \tag{2.2.1}$$

where $G\colon \mathcal{Z} \to \mathcal{X}$, $D\colon \mathcal{X} \to [0, 1]$, and

$$V_{\text{IMM}}(D, G) = \mathbb{E}_{x\sim p_r}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))]. \tag{2.2.2}$$

Remark 2.2.2. We observe that V_{IMM} is the sum of two expected-value terms. The first of these captures the idea that the discriminator wants to mark with high probability points from the real data set. The second of these terms captures that the discriminator also wants to mark with low probability points from the synthetic data set. It also captures the contrasting goal of the generator, which is to fool the discriminator into marking synthetic points with high probability.

Remark 2.2.3. This set-up is **idealised**, in that it searches for the optimal D and G over the space of *all* functions with the correct domain and co-domain.

Using this specification of the GAN game, we wish to show that the objective function is met precisely when $p_r = p_G$. If this is the case, then a GAN is successfully trained if and only if the generator distribution matches the target distribution: this is precisely what we require of GANs. To show this, we first give a result specifying the optimal discriminator, given a fixed generator.

Theorem 2.2.4 (Goodfellow et al. (2014), Proposition 1)**.** *Fix G in the IMM-GAN game. The optimal discriminator D as required by the maximisation term in* (2.2.1) *is given by*

$$D^*(x) = \frac{p_r(x)}{p_r(x) + p_G(x)}. \tag{2.2.3}$$

Proof. Maximising V_{IMM} with respect to D is equivalent to maximising

$$\begin{aligned} V_{\text{IMM}}(D, G) &= \mathbb{E}_{x\sim p_r}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \\ &= \int_{\mathcal{X}} p_r(x) \log(D(x)) \, \mathrm{d}x + \int_{\mathcal{Z}} p_z(z) \log(1 - D(G(z))) \, \mathrm{d}z \\ &= \int_{\mathcal{X}} \big[p_r(x) \log(D(x)) + p_G(x) \log(1 - D(x))\big] \, \mathrm{d}x. \end{aligned}$$

For any $(a, b) \in \mathbb{R}^2 \setminus \{(0, 0)\}$, the function $y \mapsto a \log y + b \log(1 - y)$ achieves its maximum in $[0, 1]$ at $\frac{a}{a+b}$. Since the discriminator does not need to be defined outside of the values of x for which p_r and p_G are non-zero, this concludes the proof. □

Corollary 2.2.5 (Goodfellow et al. (2014))**.** *The IMM-GAN game is equivalent to finding*

$$\min_G C(G), \tag{2.2.4}$$

where $G\colon \mathcal{Z} \to \mathcal{X}$, and

$$C(G) = \mathbb{E}_{x\sim p_r}\left[\log \frac{p_r(x)}{p_r(x) + p_G(x)}\right] + \mathbb{E}_{x\sim p_G}\left[\log \frac{p_G(x)}{p_r(x) + p_G(x)}\right]. \tag{2.2.5}$$

2.3 Divergences for GANs

To show that the IMM-GAN objective function makes sense, we need to introduce one last important element: the notion of a *divergence* between two probability distributions. This is akin to a measure of distance between two distributions: if we minimise the divergence, we also hope that the two distributions are in fact equal. Divergences come up again in the next chapter, but for now it suffices to consider two possible definitions of divergence.

Definition 2.3.1. Let μ and ν be two probability measures, and suppose μ is absolutely continuous with respect to ν. The **Kullback-Leibler (KL) divergence** from ν to μ is defined as

$$d_{\mathrm{KL}}\left(\mu\|\nu\right) = \mathbb{E}_{x\sim\mu}\left[\log\frac{\mu(x)}{\nu(x)}\right] = \mathbb{E}_{x\sim\mu}[\log\mu(x) - \log\nu(x)]. \tag{2.3.1}$$

The **Jensen-Shannon (JS) divergence** is defined as

$$d_{\mathrm{JS}}\left(\mu\|\nu\right) = \frac{1}{2}d_{\mathrm{KL}}\left(\mu\|M\right) + \frac{1}{2}d_{\mathrm{KL}}\left(\nu\|M\right), \tag{2.3.2}$$

where $M = (\mu+\nu)/2$.

Two important properties of these divergences make them useful as a notion of the difference between two distributions.[3]

Proposition 2.3.2 (Kullback and Leibler (1951), Lemma 3.1)**.** *Let μ, ν be two distributions for which the KL divergence is defined. Then $d_{\mathrm{KL}}\left(\mu\|\nu\right)$ is non-negative, and equal to 0 if and only if μ and ν are equal almost everywhere. If μ and ν are discrete probability distributions, this is equivalent to μ being equal to ν.*

Proof. This proof relies on log being concave, and $-\log$ thus being convex. Consider the case in which μ and ν are continuous probability distributions (the proof works for discrete distributions *mutatis mutandis*). Then, by Jensen's inequality:

$$\begin{aligned} d_{\mathrm{KL}}\left(\mu\|\nu\right) &= \mathbb{E}_{x\sim\mu}\left[\log\frac{\mu(x)}{\nu(x)}\right] \\ &= \int -\mu(x)\log\frac{\nu(x)}{\mu(x)}\mathrm{d}x \\ &\geq -\log\left(\int \mu(x)\cdot\frac{\nu(x)}{\mu(x)}\mathrm{d}x\right) \\ &= -\log(1) = 0. \end{aligned}$$

Hence, $d_{\mathrm{KL}}\left(\mu\|\nu\right) \geq 0$. If $\mu = \nu$ almost everywhere, then it is clear from the definition that $d_{\mathrm{KL}}\left(\mu\|\nu\right) = 0$. Moreover, since log is a *strictly* convex function, then the weak inequality is an equality only if $\mu = \nu$ almost everywhere. □

[3]The KL divergence is not a distance function as it is not symmetric - it is possible that $d_{\mathrm{KL}}\left(\mu\|\nu\right) \neq d_{\mathrm{KL}}\left(\nu\|\mu\right)$. For an example of this, see Goodfellow (2016, Figure 14). The JS divergence is a symmetrised version of the KL divergence.

Corollary 2.3.3. *Let μ, ν be two distributions for which the JS divergence is defined. Then $d_{\mathrm{JS}}(\mu\|\nu)$ is non-negative, and equal to 0 if and only if μ and ν are equal almost everywhere. If μ and ν are discrete probability distributions, this is equivalent to μ being equal to ν.*

Proof. This follows from the above proposition, and noting that the JS divergence is defined as the sum of two (non-negative) KL divergences. □

2.4 Appropriateness of Objective Function

The following theorem shows that the choice of objective function for the IMM-GAN is well-motivated.

Theorem 2.4.1 (Goodfellow et al. (2014), Theorem 1)**.** *The global minimum of the training criterion $C(G)$ is achieved if and only if $p_r = p_G$. At that point, $C(G)$ achieves the value $-\log 4$.*

Proof. For $p_r = p_G$, (2.2.3) gives us that $D^*(x) = \frac{1}{2}$, so that $C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$. To see that this is the minimal value of $C(G)$, reached only for $p_r = p_G$, observe that

$$\mathbb{E}_{x\sim p_r}[-\log 2] + \mathbb{E}_{x\sim p_G}[-\log 2] = -\log 4.$$

By subtracting this expression from $C(G) = V(D^*, G)$, we obtain:

$$\begin{aligned} C(G) &= -\log 4 + d_{\mathrm{KL}}\left(p_r \| \frac{p_{data} + p_G}{2}\right) + d_{\mathrm{KL}}\left(p_G \| \frac{p_r + p_G}{2}\right) \\ &= -\log 4 + 2 \cdot d_{\mathrm{JS}}\left(p_r \| p_G\right). \end{aligned}$$

Since the JS divergence between two distributions is always non-negative and zero only when the distributions are equal, we have that $C^* = -\log 4$ is the global minimum of $C(G)$ whose only solution is $p_r = p_G$. □

Using this proof, we may also establish that in the ideal case in which we may make updates within the function space, a broad class of convex optimisation algorithms may find this unique solution.

Theorem 2.4.2 (Adapted from Goodfellow et al. (2014), Proposition 2)**.** *The function*

$$U(p_G, D) = \mathbb{E}_{x\sim p_r}[\log D(x)] + \mathbb{E}_{x\sim p_G}[\log(1 - D(x))].$$

is convex in p_G.

Proof. We observe that only the second term depends on p_G. The proof then follows from the linearity of expectation. □

2.5 Practical Implementation of GANs

In practice we do not search over all possible functions G and D for our optima. Instead, we consider a family of parametrised functions $G(z;\theta_G)$ and $D(x;\theta_D)$ and optimise parameters $\theta_G \in \Theta_G$ and $\theta_D \in \Theta_D$. The typical class of parametrised functions we consider are **neural networks**, often abbreviated to **neural nets**.

Figure 2.1: Samples of images of bedrooms generated by a GAN trained on the LSUN dataset, taken from Goodfellow (2016).

A precise formalisation of neural nets is beyond the scope of this dissertation: the interested reader is referred to Goodfellow et al. (2016, Chapter 6). It suffices to know that neural nets have the power to approximate a broad class of functions, and are differentiable with respect to their defining parameters. The latter fact means that an objective function defined in terms of a neural net may be maximised (resp. minimised) by taking steps in the parameter space proportional to the *negative* (resp. the *positive*) of the gradient, in a process referred to as *gradient descent*.[4]

Restricting our generator and discriminator to be neural nets allows for the GAN to be implemented and trained in practice.

Algorithm 1 Minibatch stochastic gradient descent training of MM-GANs (Goodfellow et al., 2014). The gradient-based updates can use any standard gradient-based learning rule.

---

for number of training iterations **do**
  for k steps **do**
    Sample minibatch of m noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior $p_z(z)$.
    Sample minibatch of m examples $\{x^{(1)}, ..., x^{(m)}\}$ from real data distribution p_r.
    Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} \big[\log D\big(x^{(i)}\big) + \log\big(1 - D\big(G\big(z^{(i)}\big)\big)\big)\big].$$

  end for
  Sample minibatch of m noise samples $\{z^{(1)}, ..., z^{(m)}\}$ from noise prior $p_z(z)$.
  Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \log\big(1 - D\big(G\big(z^{(i)}\big)\big)\big).$$

end for

We formalise the new objective as follows:

Definition 2.5.1 (MM-GAN)**.** Let $\mathcal{Z} \subseteq \mathbb{R}^\ell$, $\mathcal{X} \subseteq \mathbb{R}^d$ be ambient data spaces, let $p_z(z)$ be a prior distribution over $\mathcal{Z}$, and let p_r be the distribution of real data points over $\mathcal{X}$. Let Θ_D and Θ_G be the spaces of possible parameters for the discriminator and generator, respectively.[5] The **minimax GAN** (MM-GAN) is the game specified by the objective

$$\min_{\theta_G \in \Theta_G} \max_{\theta_D \in \Theta_D} V_{\mathrm{MM}}(D_{\theta_D}, G_{\theta_G}), \tag{2.5.1}$$

where $G_{\theta_G} \colon \mathcal{Z} \to \mathcal{X}$, $D_{\theta_D} \colon \mathcal{X} \to [0, 1]$ belong to classes of neural nets

$$\mathcal{F} = \{G_{\theta_G} \mid \theta_G \in \Theta_G\},$$
$$\mathcal{G} = \{D_{\theta_D} \mid \theta_D \in \Theta_D\},$$

[4]The objective-maximising equivalent is also referred to as *gradient ascent*.

[5]Typically, Θ_D and Θ_G are subsets of the unit ball.

and

$$V_{\mathrm{MM}}(D_{\theta_D}, G_{\theta_G}) = \mathbb{E}_{x\sim p_r}[\log(D_{\theta_D}(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D_{\theta_D}(G_{\theta_D}(z)))]. \qquad (2.5.2)$$

2.6 Convergence Problems

Unlike its idealised counterpart, the MM-GAN objective function lacks counterparts to Theorems 2.4.1 and 2.4.2 that guarantee convergence to a unique solution such that $p_r = p_G$. This section reviews two particular problems with convergence observed when implementing this GANs in practice, and considers the theoretical explanations of their origins that have been offered. The remainder of this dissertation will focus on theoretically-motivated modifications of GANs that seek to ameliorate these problems.

2.6.1 Failure to Improve

There are two ways in which our generator may fail to improve, where the *improvement* is taken with respect to the quality of samples it produces. In the first case, though a solution may exist, the dynamics of the gradient descent training algorithm prevents the neural nets from reaching their optimal parameter values. Example 4.4.1 demonstrates this.

In the second case, which is a special case of the first failure, the gradient along which the generator must train is diminished to the point that the generator cannot usefully learn from it. This is known as the **vanishing gradient problem**, or the **saturation problem**. Goodfellow et al. (2014) claims that this problem is caused by the discriminator successfully rejecting generator samples with high confidence, so that the generator's gradient vanishes. This suggests that we ought to avoid over-training the discriminator, and instead carefully interplay discriminator and generator improvements.

2.6.2 Mode Collapse

Mode collapse is a problem that occurs when the generator learns to produce only a limited range of samples from the real data distribution. It does so by mapping several different input values $z \sim p_z$ to the same output point $G(z)$.

The name 'mode collapse' comes from the fact that, when trying to learn a multimodal distribution, the generator only outputs samples from a select number of these modes. Metz et al. (2016) demonstrates this by showing how a GAN may fail to learn a toy data distribution consisting of a mixture of 2D Gaussian distributions (as seen in Figure 2.2).

Goodfellow et al. (2014) and Metz et al. (2016) postulate that mode collapse arises from the following fact, which I shall state and prove rigorously:

Proposition 2.6.1. *Fix a continuous D in the IMM-GAN game, and let $\mathcal{X}$ be compact.*[6] *The optimal generator G as required by the outer loop of* (2.2.1) *is given*

[6]The assumption of compactness for our data space makes sense in a practical context. By the Heine-Borel Theorem, a subset of Euclidean space is compact if and only if it is closed and bounded. Representations of real data often take this form: for example, a grayscale image can be

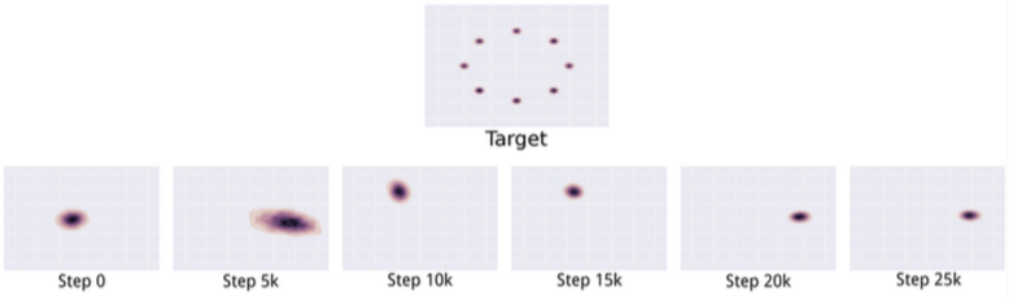

Figure 2.2: An illustration of mode collapse on a toy dataset consisting of a mixture of Gaussians in two-dimensional space. In the bottom row, we see how as the GAN is trained over time, the generator only produces a single mode at a time, cycling between different modes as the discriminator learns to reject each one. Image taken from Metz et al. (2016).

by

$$G^*(z) = s_z \quad \forall z \in \mathcal{Z}, \tag{2.6.1}$$

where $s_z \in \arg\max_{x \in \mathcal{X}} D(x)$. *In other words, the optimal generator for a fixed discriminator maps every value* $z \in \mathcal{Z}$ *to some* $x \in \mathcal{X}$ *that the discriminator believes is most likely to be real rather than fake.*

Proof. Since the discriminator is continuous with a compact domain, the set $\arg\max_{x \in \mathcal{X}} D(x)$ is non-empty. We observe that only the second term in the value function

$$\mathbb{E}_{x \sim p_r}[\log(D(x))] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

depends on G. Since log is a monotone function, we wish to choose G so as to minimise $(1 - D(G(z)))$, or equivalently, so as to maximise $D(G(z))$. From this the statement follows. □

Corollary 2.6.2. *Define* $\tilde{C}(D) := \min_G V_{\text{IMM}}(D, G)$. *Let* p_r *be any real data distribution with support on* $\mathcal{X}$ *compact. Then there exists a discriminator* D *such that* $\tilde{C}(D) = \min_G C(G)$, *but* $p_r \neq p_G$.

Proof. By the above result,

$$\tilde{C}(D) = \mathbb{E}_{x \sim p_r}[\log D(x)] + \log(1 - \max_{y \in \mathcal{X}} D(y)). \tag{2.6.2}$$

Taking D to be constantly $1/2$, we get that $\tilde{C}(D) = -\log 4 = \min_G C(G)$. However, this value for $\tilde{C}(D)$ can be attained for any G whose value is constantly $\arg\max_{y \in \mathcal{X}} D(y)$. In particular, it can be attained for a G such that $p_r \neq p_G$. □

Suppose we viewed the objective of MM-GAN as finding

$$\max_{\theta_D \in \Theta_D} \min_{\theta_G \in \Theta_G} V_{\text{MM}}(D_{\theta_D}, G_{\theta_G}).$$

given by a finite-dimensional vector of values in $[0, 1]$, and so the space of all such grayscale images will be compact. Hence, compactness is assumed here and elsewhere without loss of practical generality.

By the above proposition, this approach seemingly encourages a scenario in which the generator favours producing only one output. The problem with the GAN training algorithm, according to Goodfellow (2016), is that it does not demonstrate any preference over the **maximin** and **minimax** perspective. As emphasised by Arjovsky et al. (2017), this suggests we should seek to train the discriminator to optimality before each step of generator training.

However, this runs counter to the advice given to resolve the problem of convergence failure. As such, we need to modify our GAN design so that we may train the discriminator to optimality, avoiding the issue of mode collapse, while at the same time avoiding convergence failures.

Chapter 3

The Topology of GANs

In the previous chapter, we saw that the original GAN (Goodfellow et al., 2014) formulation suffered from problems of **convergence failure** and **mode collapse** during the training procedure. In this chapter, I review a generalisation of the GAN objective function due to Liu et al. (2017) and Zhang et al. (2017). This generalisation gives us a deeper theoretical insight into the conditions that must be satisfied for a GAN to be able to successfully reproduce samples from a distribution.

Recall that, fixing an optimal discriminator and allowing updates to the function space, finding an optimal generator is equivalent to minimising the Jensen-Shannon (JS) divergence between the generator distribution p_G and the real data distribution, p_r. The generalisation in this chapter, taking this as inspiration, shows how changing our choice of objective function makes finding the optima of that function equivalent to minimising some divergence between the two distributions.

Of course, convergence depends on our choice of distance or divergence $\rho(p_\theta, p_r)$ between these distributions. This chapter develops the argument in Arjovsky et al. (2017) that GAN training demands a distance notion d_{W} that induces a *weaker topology* than d_{JS}, in that the set of convergent sequences under d_{W} will be a superset of that under d_{JS}. I shall then show the positive theoretical results of the corresponding GAN procedure, **Wasserstein GAN** (WGAN).

3.1 Adversarial Divergences

In Chapter 2, we established that, given an optimal discriminator, we can view the IMM-GAN game as a minimisation problem for the generator. In particular, Corollary 2.2.5 showed that the IMM-GAN game was equivalent to finding

$$\min_G \left(\mathbb{E}_{x\sim p_r}\left[\log \frac{p_r(x)}{p_r(x)+p_G(x)}\right] + \mathbb{E}_{x\sim p_G}\left[\log \frac{p_G(x)}{p_r(x)+p_G(x)}\right]\right).$$

Liu et al. (2017) generalises this approach, viewing a GAN as seeking to minimise the objective function

$$p_G \mapsto \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim p_r, y\sim p_G}[f(x,y)] \tag{3.1.1}$$

for some class $\mathcal{F}$ of functions.[1] This leads to the concept of *adversarial divergence*.

[1] That our objective function is defined on a *distribution space* rather than a *parameter space* shows that this approach is 'idealised' in the sense given in the previous chapter.

Definition 3.1.1 (Modified from Liu et al. (2017), Definition 1)**.** Let $\mathcal{X}$ be an arbitrary topological space, $\mathcal{F} \subseteq [0,1]^{(\mathcal{X}^2)}$ our class of functions with domain $\mathcal{X}^2$. An **adversarial divergence** d_τ over $\mathcal{X}$ is a function

$$\begin{aligned} \text{Prob}(\mathcal{X}) \times \text{Prob}(\mathcal{X}) &\to \mathbb{R} \cup \{+\infty\} \\ (\mu, \nu) &\mapsto d_\tau(\mu \| \nu) =: \sup_{f \in \mathcal{F}} \mathbb{E}_{\mu \otimes \nu}[f]. \end{aligned}$$

Example 3.1.2 (IMM-GAN (Goodfellow et al., 2014))**.** If we set

$$\mathcal{F} = \{x, y \mapsto \log(D(x)) + \log(1 - D(y)) \mid D \in \mathcal{V}\}, \tag{3.1.2}$$

where $\mathcal{V} = [0,1]^{\mathcal{X}}$, we recover our IMM-GAN objective with optimal discriminator, $C(G)$.

Example 3.1.3 (Integral Probability Metric (Müller, 1997))**.** We derive a particularly important class of GANs when we assume that, in the definition of adversarial divergence, we can write our bivariate function f as the difference of two univariate functions.

In particular, given a choice[2] of $\mathcal{F}$, an **integral probability metric** (IPM) between two distributions is defined

$$d_{\text{IPM}}(\mu \| \nu) := \sup_{f \in \mathcal{F}} \Big(\mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{y \sim \nu}[f(y)] \Big). \tag{3.1.3}$$

Proposition 3.1.4. *Suppose that our function class $\mathcal{F}$ is such that, if $f \in \mathcal{F}$, then $-f \in \mathcal{F}$. Then $d_{\text{IPM}}(\mu \| \nu)$ is non-negative, satisfies the triangle inequality, and is symmetric.*

Proof. The proof follows easily from the properties of the supremum. □

3.2 Wasserstein GAN

This section defines the **Wasserstein GAN** (WGAN), which can be shown to arise from a particular choice of IPM. WGAN was originally developed by Arjovsky et al. (2017), after being theoretically motivated in Arjovsky and Bottou (2017). The theory has been developed further by Bousquet et al. (2017) and Lei et al. (2017).

3.2.1 Earth-Mover Distance and Total Variation Distance

We first define two notions of distance between probability distributions, both of which can be shown to be examples of IPMs.

Definition 3.2.1. Let μ, ν be probability measures on a compact metric space $(\mathcal{X}, d)$. The **Earth-Mover** (EM) or **Wasserstein-1** distance is given by

$$d_{\text{W}}(\mu \| \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \mathbb{E}_{(x,y) \sim \gamma}[\|x - y\|], \tag{3.2.1}$$

[2]Refer to Zhang et al. (2017) for an exploration of the consequences of our choice to $\mathcal{F}$ on how *useful* the consequent metric is, as well as the extent to which the empirical error bounds will *generalise* to true error bounds.

where $\Pi(\mu, \nu)$ denotes the set of all joint distributions $\gamma(x, y)$ such that, for all $A \in \mathcal{B}(\mathcal{X})$,

$$\gamma(A, \mathcal{X}) = \mu(A),$$
$$\gamma(\mathcal{X}, A) = \nu(A).$$

Intuitively, $\gamma(x, y)$ indicates how much 'mass' must be transported from x to y in order to transform the distribution μ into the distribution ν.

The following equivalent formula is more tractable when finding minima with respect to the Wasserstein distance.

Theorem 3.2.2 (The Kantorovich-Rubinstein Duality)**.** *Let $(\mathcal{X}, d)$ be a compact metric space, and let $\mathrm{Lip}_1(\mathcal{X})$ be the set of functions $f\colon \mathcal{X} \to \mathbb{R}$ such that*

$$\|f\|_{\mathrm{L}} := \sup \left\{ \frac{|f(x) - f(y)|}{d(x, y)} \;\middle|\; x, y \in \mathcal{X}, x \neq y \right\} \leq 1.$$

Then $f \in \mathrm{Lip}_1(\mathcal{X})$ is Lebesgue integrable with respect to any probability measure on $\mathcal{X}$, and

$$d_{\mathrm{W}}(\mu \| \nu) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} (\mathbb{E}_{x \sim \mu}[f(x)] - \mathbb{E}_{x \sim \nu}[f(x)]). \tag{3.2.2}$$

Proof. The result is a standard one in optimal transport theory. See, e.g., Villani (2008, Theorem 5.10) for a proof. An alternate proof can be found in Edwards (2011, Theorem 4.1). The presentation here is an adaptation of the latter approach, as given by Basso (2015, Theorem 1.3).

Since f is 1-Lipschitz, we have for some $x_0 \in \mathcal{X}$ that

$$|f(x)| \leq |f(x_0)| + d(x, x_0). \tag{3.2.3}$$

Since $\mathcal{X}$ is compact, for any probability measure $\mathbb{P}$ on $\mathcal{X}$ we can integrate both sides of (3.2.3) with respect to $\mathbb{P}$ to obtain

$$\int_{\mathcal{X}} |f(x)| \, \mathrm{d}\mathbb{P} \leq \int_{\mathcal{X}} (|f(x_0)| + d(x, x_0)) \, \mathrm{d}\mathbb{P} \leq +\infty.$$

Now let $\mathcal{B}^\infty(\mathcal{X})$ be the set of all bounded Borel-measurable functions $f\colon \mathcal{X} \to \mathbb{R}$. For $f, g\colon \mathcal{X} \to \mathbb{R}$, we define $(f \oplus g)\colon \mathcal{X}^2 \to \mathbb{R}$ by

$$(f \oplus g)(x, y) := f(x) + g(y).$$

Observe that since μ, ν are probability measures on a compact metric space, they are (bounded non-negative) Radon measures. Further, $d\colon \mathcal{X}^2 \to \mathbb{R}$ is continuous, and hence lower semicontinuous, as a distance metric. By Corollary 3.2 of Edwards (2011), we have that

$$d_{\mathrm{W}}(\mu \| \nu) = \sup \left\{ \int_{\mathcal{X}} f \, \mathrm{d}\mu + \int_{\mathcal{X}} g \, \mathrm{d}\nu \mid f, g \in \mathcal{B}^\infty(\mathcal{X}), (f \oplus g) \leq d \right\}.$$

Fix $\varepsilon > 0$. By the Approximation Lemma, there exists $f, g \in \mathcal{B}^\infty(\mathcal{X})$ with $(f \oplus g) \leq d$ such that

$$d_{\mathrm{W}}(\mu \| \nu) - \varepsilon \leq \int_{\mathcal{X}} f \, \mathrm{d}\mu + \int_{\mathcal{X}} g \, \mathrm{d}\nu.$$

Now define $k\colon \mathcal{X} \to \mathbb{R}$ by $k(x) := \inf_{y\in\mathcal{X}} (d(x,y) - g(y))$. Since g is bounded, k is well-defined. Then, for $x, x' \in \mathcal{X}$, we have that

$$\begin{aligned}
|k(x) - k(x')| &= |\inf_{y\in\mathcal{X}} (d(x,y) - g(y)) - \inf_{y\in\mathcal{X}} (d(x',y) - g(y))| \\
&\leq \sup_{y\in\mathcal{X}} |d(x,y) - d(x',y)| \\
&\leq d(x,x').
\end{aligned}$$

Hence $k \in \mathrm{Lip}_1(\mathcal{X})$. Note further that, for all $x \in \mathcal{X}$,

$$f(x) \leq k(x) \leq d(x,x) - g(x) = -g(x),$$

so $f \leq k$ and $g \leq -k$. Now let $\gamma \in \Pi(\mu,\nu)$. We then get

$$\begin{aligned}
d_{\mathrm{W}}(\mu\|\nu) - \varepsilon &\leq \int_{\mathcal{X}} f \,\mathrm{d}\mu + \int_{\mathcal{X}} g \,\mathrm{d}\nu \\
&\leq \int_{\mathcal{X}} k \,\mathrm{d}\mu - \int_{\mathcal{X}} k \,\mathrm{d}\nu \\
&\leq \sup\left\{ \int_{\mathcal{X}} f \,\mathrm{d}\mu - \int_{\mathcal{X}} f \,\mathrm{d}\nu \mid f \in \mathrm{Lip}_1(\mathcal{X}) \right\} \\
&\leq \sup\left\{ \int_{\mathcal{X}\times\mathcal{X}} (f \oplus -f) \,\mathrm{d}\gamma \mid f \in \mathrm{Lip}_1(\mathcal{X}) \right\} \\
&\leq \int_{\mathcal{X}\times\mathcal{X}} d(x,y) \,\mathrm{d}\gamma(\mathrm{d}x, \mathrm{d}y).
\end{aligned}$$

Letting $\varepsilon \to 0$, we get the desired equality. □

Corollary 3.2.3. *The EM distance is an IPM, so long as the domain $\mathcal{X}$ of our function class $\mathcal{F}$ is a compact metric space.*

The following corollary tells us that it makes sense to describe the EM distance as a *distance*, and to talk of it inducing a *topology*.

Corollary 3.2.4 (Basso (2015), Corollary 1.4)**.** *Let $(\mathcal{X}, d)$ be a compact metric space. Then d_W defines a metric on* $\mathrm{Prob}(\mathcal{X})$.

Proof. That d_{W} is symmetric and non-negative is clear from (3.2.1), and that it obeys the triangle inequality is clear from (3.2.2). Hence it remains to show that for probability measures μ, ν on a compact metric space $(\mathcal{X}, d)$ that if $d_{\mathrm{W}}(\mu\|\nu) = 0$, then $\mu = \nu$.

Let F be a closed subset of $\mathcal{X}$. For each integer $k \geq 1$, we define $f_k\colon \mathcal{X} \to \mathbb{R}$ by $f_k(x) := 1 \wedge (k \cdot \mathrm{dist}(x, F))$. Then it follows that, for each integer k, $f_k/k \in \mathrm{Lip}_1(\mathcal{X})$. Furthermore, since $d_{\mathrm{W}}(\mu\|\nu) = 0$, (3.2.2) gives us that, for all $k \geq 1$,

$$\frac{1}{k} \int_{\mathcal{X}} f_k \,\mathrm{d}\mu = \frac{1}{k} \int_{\mathcal{X}} f_k \,\mathrm{d}\nu. \tag{3.2.4}$$

Observe that $(f_k)_{k\geq 1}$ is a non-negative sequence of functions converging monotonically to the indicator function on $\mathcal{X} \setminus F$, an open set. Hence, by the Monotone Convergence Theorem and (3.2.4), it follows that $\mu(\mathcal{X} \setminus F) = \nu(\mathcal{X} \setminus F)$. Since open subsets of $\mathcal{X}$ generate $\mathcal{B}(\mathcal{X})$, it follows by Dynkin's Lemma that $\mu = \nu$. □

We obtain similar results for the **Total Variation distance** between two distributions.

Definition 3.2.5. The **Total Variation** (TV) distance between μ and ν is defined

$$d_{\mathrm{TV}}(\mu\|\nu) = \sup_{A\in\mathcal{B}(\mathcal{X})} |\mu(A) - \nu(A)|. \tag{3.2.5}$$

Proposition 3.2.6. *The TV distance is an IPM, where $\mathcal{F}$ is the set of of all measurable functions bounded between -1 and 1.*

Proof. See Müller (1997). □

Corollary 3.2.7. *The TV distance defines a metric on* $\mathrm{Prob}(\mathcal{X})$.

Proof. By Corollary 3.1.4, it suffices to prove that $d_{\mathrm{TV}}(\mu\|\nu) = 0$ implies $\mu = \nu$. But this is given from the definition of d_{TV} as a supremum: in particular, μ and ν are equal on all Borel measurable subsets, and so must be equal. □

3.2.2 The Weakness of the Wasserstein Distance

We now seek to show that, in some rigorous sense, minimising Wasserstein distance is a more suitable framework for GAN training than minimising the Jensen-Shannon divergence. The adversarial divergence framework enables us to view the objective of GANs as the minimisation of some divergence between our generator distribution p_G and our target distribution p_r. Moreover, the framework can be used to consider the *training* of a GAN as the convergence of p_G to p_r with respect to the given divergence.

With distinct definitions of convergence come distinct induced *topologies*.[3] The idea of convergence gives rise to an idea of a *topology* induced by the divergence. Notably, we ought to seek some divergence that give us a *weak* topology, in that the convergence of p_G to p_r with respect to other divergences implies convergence with respect to our ideal divergence. Using the language of functional analysis, it can be shown that the Wasserstein distance meets this desideratum.[4]

Let $\mathcal{X}$ be a compact set. Taking the sup-norm $\|f\|_\infty = \max_{x\in\mathcal{X}}|f(x)|$, the space $(\mathcal{C}_b(\mathcal{X}), \|\cdot\|_\infty)$ is a normed vector space. We can then define the dual normed space $(\mathcal{C}_b(\mathcal{X})^*, \|\cdot\|)$, where we take

$$\begin{aligned}\mathcal{C}_b(\mathcal{X})^* &:= \{\phi\colon \mathcal{C}_b(\mathcal{X}) \to \mathbb{R} \mid \phi \text{ is linear and continuous.}\}\\ \|\phi\| &:= \sup_{f\in\mathcal{C}_b(\mathcal{X}),\|f\|_\infty\leq 1} |\phi(f)|.\end{aligned}$$

Consider the mapping

$$\begin{aligned}\Phi\colon (\mathrm{Prob}(\mathcal{X}), d_{\mathrm{TV}}) &\to (\mathcal{C}_b(\mathcal{X})^*, \|\cdot\|)\\ \Phi(\mu)(f) &:= \mathbb{E}_{x\sim\mu}[f(x)].\end{aligned}$$

By linearity of expectation, this function indeed maps to the appropriate dual space and so is well-defined. Therefore, by the Riesz-Markov-Kakutani representation theorem (Kakutani, 1941, Theorem 10), Φ is an isometric immersion. This

[3]Not all divergences define metrics (e.g., the KL and reverse-KL divergence). As a result, a given adversarial divergence may not give us a topology in a strictly formal sense.

[4]This argument can be found in Arjovsky et al. (2017, Appendix A).

allows us to regard convergence in TV distance and convergence with respect to $\|\cdot\|$ as essentially equivalent. This is unfortunate for the TV distance: convergence with respect to the latter norm in $\mathcal{C}_b(\mathcal{X})^*$ is regarded as 'strong' convergence,[5] in effect limiting the capacity of a TV-based GAN to train towards a variety of real distributions.

By contrast, $\mathcal{C}_b(\mathcal{X})^*$ also comes equipped with a much *weaker* topology.

Definition 3.2.8 (e.g., Liu et al. (2017), Definitions 7-8)**.** Let $\mathcal{X}$ be a compact metric space. The **weak* topology** for $\text{Prob}(\mathcal{X})$ is the coarsest topology on $\text{Prob}(\mathcal{X})$ such that

$$\{\mu \mapsto \mathbb{E}_\mu[f] \mid f \in \mathcal{C}_b(\mathcal{X})\}$$

is a set of continuous linear functions on $\text{Prob}(\mathcal{X})$.

Moreover, we say that a sequence $(\mu_n) \subseteq \text{Prob}(\mathcal{X})$ **weakly converges** to a measure $\mu \in \text{Prob}(\mathcal{X})$ if, for all $f \in \mathcal{C}_b(\mathcal{X})$,

$$\mathbb{E}_{\mu_n}[f] \to \mathbb{E}_\mu[f] \text{ as } n \to \infty,$$

or equivalently, if $\mu_n \to \mu$ in the weak* topology.

If we can show that the Wasserstein distance captures the notion of weak* convergence, then we may claim that the WGAN gives us a more suitable choice of objective function than any other GAN. We can go further than this by providing a hierarchy of divergences with respect to their convergence.

Theorem 3.2.9 (Arjovsky et al. (2017), Theorem 2)**.** *Let $\mathcal{X}$ be compact, and let $\mu, (\mu_n) \subseteq \text{Prob}(\mathcal{X})$. Then, considering all limits as $n \to \infty$,*

1. $d_{\text{TV}}(\mu_n \| \mu) \to 0$ *if and only if* $d_{\text{JS}}(\mu_n \| \mu) \to 0$.
2. $d_{\text{W}}(\mu_n \| \mu) \to 0$ *if and only if* $\mu_n \xrightarrow{d} \mu$, *where* $\xrightarrow{d}$ *denotes convergence in distribution.*
3. *If either* $d_{\text{KL}}(\mu \| \mu_n) \to 0$ *or* $d_{\text{KL}}(\mu_n \| \mu) \to 0$, *then* $d_{\text{JS}}(\mu_n \| \mu) \to 0$.
4. *If* $d_{\text{JS}}(\mu_n \| \mu) \to 0$, *then* $d_{\text{W}}(\mu_n \| \mu) \to 0$.

Proof.

1. See Appendix A.
2. This comes from the standard result that d_{W} gives a metric for the weak* topology of $(\mathcal{C}_b(\mathcal{X}), \|\cdot\|_\infty)$ on $\text{Prob}(\mathcal{X})$, and by definition, this is the topology of convergence in distribution. For a proof, see Villani (2008, Theorem 6.9).
3. By Pinsker's Inequality (Cesa-Bianchi and Lugosi, 2006, Section A.2, p. 371), either case gives us one of

$$d_{\text{TV}}(\mu_n \| \mu) \leq \sqrt{\frac{1}{2} d_{\text{KL}}(\mu \| \mu_n)} \to 0,$$
$$d_{\text{TV}}(\mu_n \| \mu) \leq \sqrt{\frac{1}{2} d_{\text{KL}}(\mu_n \| \mu)} \to 0.$$

[5] It is referred to as such in the standard literature. For example, see Kreyszig (1978, Definition 4.9-4).

4. This comes from the fact, as argued above, that d_{TV} and d_{W} give the strong and weak* topologies on the dual of $(\mathcal{C}_b(\mathcal{X}), \|\cdot\|_\infty)$ when restricted to the space of probability measures on $\mathcal{X}$.

□

Finally, we observe how even a simple sequence of probability distributions converges under d_{W} but not under d_{JS} or d_{KL}. This serves as a witness to Theorem 3.2.9.

Example 3.2.10 (Arjovsky et al. (2017), Example 1)**.** *Let $Z \sim U[0,1]$ be uniformly distributed over the unit interval. Let $\mathbb{P}_0$ be the distribution of $(0, Z) \in \mathbb{R}^2$. Now let $g_\theta(z) = (\theta, z)$, with $\theta \in \mathbb{R}$ and $\mathbb{P}_\theta$ the distribution for $g_\theta(Z)$. In this case:*

- $d_{\mathrm{W}}(\mathbb{P}_0 \| \mathbb{P}_\theta) = |\theta|$,
- $d_{\mathrm{JS}}(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} \log 2 & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0, \end{cases}$
- *and* $d_{\mathrm{KL}}(\mathbb{P}_\theta \| \mathbb{P}_0) = d_{\mathrm{KL}}(\mathbb{P}_0 \| \mathbb{P}_\theta) = \begin{cases} +\infty & \text{if } \theta \neq 0, \\ 0 & \text{if } \theta = 0. \end{cases}$

Hence, when $\theta_n \to 0$, the sequence $(P_{\theta_n})_{n \in \mathbb{N}}$ converges to $\mathbb{P}_0$ only under the Wasserstein distance.

3.2.3 On the Viability of WGAN

In the above example, the JS divergence fails to give us continuous mapping $\theta \mapsto \mathbb{P}_\theta$, a desirable property. The next theorem shows us that, under mild assumptions, $d_{\mathrm{W}}(\mu \| \mu_\theta)$ is a continuous loss function on θ.

Theorem 3.2.11 (Arjovsky et al. (2017), Theorem 1)**.** *Let $\mathcal{X}$ be compact, and let $\mu \in \mathrm{Prob}(\mathcal{X})$. Let Z be a random variable over another space $\mathcal{Z}$. Let $g\colon \mathcal{Z} \times \mathbb{R}^d \to \mathcal{X}$ be a function, denoted $g_\theta(z)$. Let μ_θ denote the distribution of $g_\theta(Z)$. Then,*

1. *If g is continuous in θ, so is $d_{\mathrm{W}}(\mu \| \mu_\theta)$,*
2. *If g is locally Lipschitz and there are local Lipschitz constants $L(\theta, z)$ such that*

$$\mathbb{E}_{z \sim \mu_\theta}[L(\theta, z)] < +\infty,$$

   then $d_{\mathrm{W}}(\mu \| \mu_\theta)$ is continuous everywhere, and differentiable almost everywhere.

3. *Statements 1-2 are false for d_{JS} and the two KL divergences.*

Proof.

1. Let γ be the distribution of $(g_\theta(Z), g_{\theta'}(Z))$, so that $\gamma \in \Pi(\mu_\theta, \mu_{\theta'})$. Then (3.2.1) gives us

$$\begin{aligned} d_{\mathrm{W}}(\mu_\theta \| \mu_{\theta'}) &\leq \int_{\mathcal{X} \times \mathcal{X}} \|x - y\| \, \mathrm{d}\gamma \\ &= \mathbb{E}_{(x,y) \sim \gamma}[\|x - y\|] \\ &= \mathbb{E}_{\mathcal{Z}}[\|g_\theta(Z) - g_{\theta'}(Z)\|]. \end{aligned}$$

Since g is continuous in θ, we have that $g_\theta(z) \xrightarrow{\theta \to \theta'} g_{\theta'}(z)$. Hence

$$\|g_\theta - g_{\theta'}\| \to 0$$

point-wise in z. Since $\mathcal{X}$ is compact, there exists a positive constant M, independent of θ and z, such that for all θ, z, we have $\|g_\theta(z) - g_{\theta'}(z)\| \leq M$. By the Bounded Convergence Theorem,

$$\begin{aligned} |d_{\mathrm{W}}(\mu \| \mu_\theta) - d_{\mathrm{W}}(\mu \| \mu_{\theta'})| &\leq d_{\mathrm{W}}(\mu_\theta \| \mu_{\theta'}) \\ &\leq \mathbb{E}_{\mathcal{Z}}[\|g_\theta(z) - g_{\theta'}(z)\|] \\ &\xrightarrow{\theta \to \theta'} 0. \end{aligned}$$

The result follows.

2. Take g to be locally Lipschitz. This means that, fixing (θ, z), there exists a constant $L(\theta, z)$ and open set U such that $(\theta, z) \in U$, and for every $(\theta', z') \in U$,

$$\|g_\theta(z) - g_{\theta'}(z')\| \leq L(\theta, z)(\|\theta - \theta'\| + \|z - z'\|). \tag{3.2.6}$$

Fix $z' = z$. Taking the expectation in (3.2.6), we have for all $(\theta', z) \in U$ that

$$\mathbb{E}_{\mathcal{Z}}[\|g_\theta(z) - g_{\theta'}(z)\|] \leq \|\theta - \theta'\| \cdot \mathbb{E}_{\mathcal{Z}}[L(\theta, z)].$$

Define $U_\theta := \{\theta' \mid (\theta', z) \in U\}$. Since U is open, U_θ is also open. Hence, by hypothesis, we may define $L(\theta) := \mathbb{E}_{\mathcal{Z}}[L(\theta, z)]$ and get for all $\theta' \in U_\theta$ that

$$|d_{\mathrm{W}}(\mu \| \mu_\theta) - d_{\mathrm{W}}(\mu \| \mu_{\theta'})| \leq d_{\mathrm{W}}(\mu_\theta \| \mu_{\theta'}) \leq L(\theta) \cdot \|\theta - \theta'\|.$$

Hence $d_{\mathrm{W}}(\mu \| \mu_\theta)$ is locally Lipschitz. Therefore, $d_{\mathrm{W}}(\mu \| \mu_\theta)$ is everywhere continuous, and, by Rademacher's Theorem (Federer, 2014, Theorem 3.1.6), differentiable almost everywhere.

3. Observe that Example 3.2.10 serves as the required counterexample.

□

Of course, in practice our generator functions will be neural nets. The following corollary shows that the previous result holds if we restrict our attention to these kinds of functions.

Corollary 3.2.12 (Arjovsky et al. (2017), Corollary 1)**.** *Let g_θ be any feedforward neural net*[6] *parametrised by θ, and p_z a prior over z such that $\mathbb{E}_{z \sim p_z}[\|z\|] < \infty$. Then the assumptions of Theorem 3.2.11 are satisfied, and so $d_{\mathrm{W}}(\mu \| \mu_\theta)$ is continuous everywhere and differentiable almost everywhere.*

Proof. Since g is a continuously differentiable function in (θ, z), for any fixed (θ, z), we have for all $\varepsilon > 0$ that $L(\theta, z) \leq \|\nabla_{\theta, x} g_\theta(z)\| + \varepsilon$ is an acceptable local Lipschitz constant. It therefore remains to show that

$$\mathbb{E}_{z \sim p_z}[\|\nabla_{\theta, x} g_\theta(z)\|] < +\infty. \tag{3.2.7}$$

This part of the proof is omitted. Refer to Arjovsky et al. (2017, Appendix C) for the technical details. □

[6] In other words, a function composed by affine transformations and pointwise nonlinearities which are smooth Lipschitz functions.

3.2.4 The WGAN Procedure

Evaluating $d_{\mathrm{W}}(p_r \| p_\theta)$, where p_θ is the distribution of our generator g_θ, is often intractable. A more tractable approach, justified by the Kantorovich-Rubinstein Duality, would be to solve the problem

$$\max_{w \in \mathcal{W}} \big(\mathbb{E}_{x \sim p_r}[f_w(x)] - \mathbb{E}_{z \sim p_z(z)}[f_w(g_\theta(z))] \big), \tag{3.2.8}$$

where $\{f_w\}_{w \in \mathcal{W}}$ is a set of parametrised functions that are all K-Lipschitz for some K. If the supremum in (3.2.2) is attained for some $w \in \mathcal{W}$, this process would yield a calculation of $d_{\mathrm{W}}(p_r \| p_\theta)$ up to a multiplicative constant K.

To minimise $d_{\mathrm{W}}(p_r \| p_\theta)$ with respect to θ, we could consider differentiating $d_{\mathrm{W}}(p_r \| p_\theta)$ (up to a constant) by using back-propagation through equation (3.2.2) via estimating $\mathbb{E}_{z \sim p_z(z)}[\nabla_\theta f_w(g_\theta(z))]$. The following theorem shows that this process is principled under the assumptions of the previous results.

Theorem 3.2.13 (Arjovsky et al. (2017), Theorem 3)**.** *Let $\mathcal{X}$ be compact, and let $p_r \in \mathrm{Prob}(\mathcal{X})$. Let p_θ be the distribution of $g_\theta(Z)$ with Z a random variable with density p_z and g_θ a function satisfying the assumptions of Theorem 3.2.11. Then, there is a solution $f : \mathcal{X} \to \mathbb{R}$ to the problem*

$$\max_{f \in \mathrm{Lip}_1(\mathcal{X})} \big(\mathbb{E}_{x \sim p_r}[f(x)] - \mathbb{E}_{x \sim p_\theta}[f(x)] \big)$$

and we have

$$\nabla_\theta d_{\mathrm{W}}(p_r \| p_\theta) = -\mathbb{E}_{z \sim p_z(z)}[\nabla_\theta f(g_\theta(z))]$$

when both terms are well-defined.

Proof. The proof can be found in Appendix A. □

To roughly approximate finding the function f that solves (3.2.2), Arjovsky et al. (2017) train a neural network parametrised with weights w lying in a compact space $\mathcal{W}$, and then perform back-propagation through $\mathbb{E}_{z \sim p_z(z)}[\nabla_\theta f_w(g_\theta(z))]$.[7]

Here, we refer to f_w as our *critic*, just as the original GAN had a *discriminator* D. When $\mathcal{W}$ is compact,[8] all the functions f_w will be K-Lipschitz for some K depending only on $\mathcal{W}$ and not the individual weights w, allowing us approximate (3.2.2) up to an irrelevant scaling factor.

3.2.5 Does WGAN Resolve Convergence Problems?

Mode Collapse

Arjovsky et al. recommend that WGAN critics be trained to optimality. In so doing, they claim, one avoids the phenomenon of **mode collapse**. The argument for this relies on the Goodfellow-Metz explanation for mode collapse. The reasoning is as

[7] Refer to Algorithm 1 in Arjovsky et al. (2017) for the precise formulation of the WGAN algorithm.

[8] This is enforced in the WGAN algorithm given by Arjovsky et al. (2017) by clipping the weights within a fixed hypercube, say $\mathcal{W} = [-0.01, 0.01]^l$, after each gradient update. See Gulrajani et al. (2017) for an alternate approach to enforcing the Lipschitz constraint by adding a 'gradient penalty' term to the discriminator loss function.

follows: the hypothesis is that mode collapse comes from the fact (Proposition 2.6.1) that the optimal generator for a given discriminator is a map to the points for which the discriminator assigns the highest probability, so that each generator update step in training the MM-GAN is a partial collapse towards this function. Therefore, if the discriminator is close enough to optimality, the data points to which it will assign the highest probability will be precisely those that arise from the real data distribution p_r, and so an optimal generator function given this discriminator will be a delta function valued 1 on all such data points.

It has also been observed empirically that WGANs avoid mode collapse in cases where the MM-GANs do not (Arjovsky et al., 2017; Fedus et al., 2017; Lucic et al., 2017).

Failure to Converge

According to Arjovsky et al. (2017), the fact that the EM distance is continuous and differentiable a.e. means that, unlike the MM-GAN, we can train the critic to optimality without worrying about the convergence of the generator distribution.

Consider, for instance, the target and generator distributions given in Example 3.2.10. Unless the generator distribution has already matched the target distribution, the JS distance between the two is constantly $\log 2$. This distance, as a result, gives no meaningful gradient for the generator to use for training, provided that the discriminator is sufficiently optimal so as to give an accurate estimate of the JS distance.[9] By contrast, the WGAN critic converges to a piecewise linear function through the constraint of its weights by clipping. In this sense, we see how the WGAN can tackle the issue of saturation.

[9] Arjovsky and Bottou (2017, Theorem 2.4) shows that this issue of vanishing gradients for the MM-GAN generator can be found to occur quite generally.

Chapter 4

A Game-Theoretic Analysis of GANs

4.1 Many Paths to Equilibrium

Throughout the previous chapter, I considered the training of GANs as a minimisation of an *adversarial divergence*, itself the supremum over some loss function. In doing so, I took the perspective of GANs as optimising a generator, given a discriminator that is already optimal.

While such an approach does enjoy certain theoretical and empirical successes,[1] there are also two potential drawbacks.

Firstly, it may be computationally impractical to train a discriminator to optimality at each step. An approach that trains the generator and discriminator simultaneously, or trains the discriminator for k iterations between each generator training step may simply converge to the solution more efficiently. When such an approach is taken with the neural net parameters trained via gradient descent, these approaches are referred to as **simultaneous** and **alternating gradient descent** (SimGD and AGD), respectively.

Secondly, as we have seen, certain GAN formulations do not perform well with optimal discriminators: the gradient along which the generator is being optimised collapses, meaning it converges to its optimum far more slowly. This was observed in an informal setting for MM-GANs in the original GAN paper by Goodfellow et al. (2014).

4.1.1 Two Generalisations

In this chapter, I will investigate the consequences of relaxing two constraints implicitly imposed by the *adversarial divergence* view of training GANs.

1. We will now consider the cases in which the generator and discriminator are trained via simultaneous or alternating gradient descent. The former approach has been modelled in Mescheder et al. (2017). The latter was first proposed by Goodfellow et al. (2014) (Algorithm 1 in this dissertation), with the number k of discriminator training steps between each generator training step treated

[1] For instance, in Wasserstein GAN (Arjovsky et al., 2017; Arjovsky and Bottou, 2017).

as an algorithm hyperparameter to be carefully chosen.[2]

2. We will allow for the discriminator and generator to be trained on objective functions whose absolute values differ. In particular, we now consider the discriminator and generator to seek to minimise **loss functions** J^D and J^G, respectively. Hence, unless $J^D = -J^G$ as assumed in the previous chapters, we can no longer express the GAN objective by a single value function V.

For a simple example of a GAN that cannot be expressed by a single value function, consider the following example.

Definition 4.1.1 (Goodfellow et al. (2014))**.** Let $\mathcal{Z} \subseteq \mathbb{R}^\ell$, $\mathcal{X} \subseteq \mathbb{R}^d$ be ambient data spaces, let p_z be a prior distribution over $\mathcal{Z}$, and let p_r be the distribution of real data points over $\mathcal{X}$. The **non-saturating GAN** (NS-GAN) is the game specified by the minimisation of loss functions

$$J^D_{\text{NS}} := \mathbb{E}_{x\sim p_r}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))],$$
$$J^G_{\text{NS}} := -\mathbb{E}_{z\sim p_z}[\log D(G(z))].$$

where $G\colon \mathcal{Z} \to \mathcal{X}$, $D\colon \mathcal{X} \to [0, 1]$.

Such a game has been observed to enjoy empirical success: in particular, Fedus et al. (2017) show that on a variety of datasets, NS-GAN produces samples of a comparable quality to WGAN and its variants, while being easier to train.[3]

4.2 A Brief Review of Game Theory

Since the NS-GAN is not expressible by a single value function, our theoretical results about adversarial divergence minimisation no longer apply. To analyse GAN examples like NS-GAN, we require a game-theoretic framework. In doing so, we can provide answers to the following questions:

> Do solutions to the GAN game exist? What is the nature of such solutions?
>
> Are there training methods for the discriminator and generator that allow us to converge to such solutions?

4.2.1 Two-Player Games

We begin with the definition of the type of game we are interested in. Unless stated otherwise, the definitions and results are adapted from Osborne and Rubinstein (1994, Chapters 2-3).

[2]For a comparison between AGD and training the discriminator to optimality on a toy example, refer to Li et al. (2017).

[3]Visual inspection is currently one of the most prominent methods of evaluation within the GAN literature. In general, finding a suitable quantitative evaluation model for GANs is one of the largest open problems in the field (Goodfellow, 2016, p. 42). Currently, the Inception Score (Salimans et al., 2016) and the Fréchet Inception Distance (Ramsauer et al., 2017) are amongst the most popular performance metrics available.

Definition 4.2.1. A **strategic two-player game** $\langle (A_i), (u_i) \rangle_{i=1,2}$ consists of, for players $i = 1, 2$

- a nonempty set A_i (the set of **actions** available to player i)
- a **payoff or utility function** $u_i \colon A \to \mathbb{R}$, where $A = A_1 \times A_2$.

If the set A_i of actions of both players is finite, then the game is **finite**.

This game is referred to as *strategic* as each player chooses their plan of action once and for all, with these choices being made simultaneously. Given such a game, the most commonly used solution concept is that of a **Nash equilibrium**. The Nash equilibrium captures the idea that each player holds the correct expectation about the other players' behaviour and acts rationally; in a Nash equilibrium, neither player can gain by deviating from their strategy.

Definition 4.2.2. Let a_{-i} denote the strategy of player 1 for $i = 2$, and player 2 for $i = 1$. A **Nash equilibrium of a strategic two-player game** $\langle (A_i), (u_i) \rangle_{i=1,2}$ is a profile $a^* \in A$ of actions with the property that for $i = 1, 2$ we have

$$u_i(a^*_{-i}, a^*_i) \geq u_i(a^*_{-i}, a_i) \text{ for all } a_i \in A_i.$$

4.2.2 Minimax Games

Definition 4.2.3. A strategic two-player game $\langle (A_i), (u_i) \rangle_{i=1,2}$ is **zero-sum** if $u_1 = -u_2$.

Remark 4.2.4. With an appropriate choice of *minimax theorem*,[4] we can show that an optimal solution to our adversarial divergence objective function exists, and that it coincides with a Nash equilibrium of a zero-sum game.

Unfortunately, such results do not apply to GANs. Typically, a minimax theorem requires the A_i to be compact and the value function $V(\cdot, \cdot)$ to be convex in the first argument and concave in the second. In general, the value function will fail to have this property due to a lack of expressivity in the discriminator-generator function space. Even if we allow the discriminator and generator to range over a broader class of functions, the associated function spaces will no longer be compact. We see this failure in the mode collapse hypothesis, in which the minimax and maximin solutions are distinct.

4.2.3 Mixed Extensions to Games

Given that the GAN game consists in practice of making updates to the parameter space rather than the function space, is there any way we can generalise the notion of a Nash equilibrium to recover the guarantee of the existence of such an equilibrium for the game? The answer is a qualified yes.

Our required generalisation is that of a *mixed strategy Nash equilibrium*: this will be a steady state of the game in which the players' choices are not deterministic, but instead determined probabilistically according to some distribution.

[4] See, e.g., Fan (1953, Theorem 2).

In particular, we denote by $\Delta(A_i)$ the set of probability distributions over A_i and refer to a member of $\Delta(A_i)$ as a **mixed strategy**[5] of player i. A member of A_i is referred to as a **pure strategy**. For any finite set X and $\delta \in \Delta(X)$ we denote by $\delta(x)$ the probability that δ assigns to $x \in X$, and define the **support** of δ to be the set of elements $x \in X$ such that $\delta(x) > 0$. Given a profile (α_1, α_2) of mixed strategies, we have a probability distribution over the set A. If each A_i is finite, then the probability of the action profile $a = (a_1, a_2)$ is $\alpha_1(a_1) \cdot \alpha_2(a_2)$, and player i's evaluation of (α_1, α_2) is $\sum_{a \in A}(\alpha_1(a_1) \cdot \alpha_2(a_2)) \cdot u_i(a)$. This leads to the following definition.

Definition 4.2.5. The **mixed extension** of the strategic two-player game $\langle (A_i), (u_i) \rangle_{i=1,2}$ is the strategic two-player game $\langle (\Delta(A_i)), (U_i) \rangle_{i=1,2}$ in which $\Delta(A_i)$ is the set of probability distributions over A_i, and $U_i \colon \Delta(A_1) \times \Delta(A_2) \to \mathbb{R}$ assigns to each $\alpha \in \Delta(A_1) \times \Delta(A_2)$ the expected value under u_i of the probability distribution over A induced by α, so that

$$U_i(\alpha) = \sum_{a \in A}(\alpha_1(a_1) \cdot \alpha_2(a_2)) \cdot u_i(a)$$

From this definition, it follows that U_i is multilinear: for any mixed strategy profile α, any mixed strategies β_i and γ_i of player i, and any number $\lambda \in [0, 1]$, that

$$U_i(\alpha_{-i}, \lambda\beta_i + (1 - \lambda)\gamma_i) = \lambda U_i(\alpha_{-i}, \beta_i) + (1 - \lambda)U_i(\alpha_{-i}, \gamma_i).$$

Moreover, when each A_i is finite, we have

$$U_i(\alpha) = \sum_{a_i \in A_i} \alpha_i(a_i) U_i(\alpha_{-i}, e(a_i))$$

for any mixed strategy profile α, where $e(a_i)$ is the degenerate mixed strategy of player i corresponding to the pure strategy of just choosing a_i.

Definition 4.2.6. A **mixed strategy Nash equilibrium of a strategic two-player game** is a Nash equilibrium of its mixed extension.

It is plain to see, since each U_i is multilinear, and any mixed strategy exclusively employing degenerate distributions can be readily identified with a strategy in A, that the set of Nash equilibria of a strategic game is a subset of its set of mixed strategy Nash equilibria.

In a well-known result, Nash (1950) showed that a mixed strategy Nash equilibrium is guaranteed to exist if the action spaces A_1, A_2 are finite. Of course, this condition fails to hold for the generator and discriminator of a GAN, whose strategy spaces are parameter spaces in $\mathbb{R}^n, \mathbb{R}^m$, respectively. However, we may restrict the parameter spaces such that the GAN game is a **continuous game**, where A_1, A_2 are non-empty compact metric spaces and u_1, u_2 are continuous functions on A. We then get a result that applies to the GAN game.

Theorem 4.2.7 (Glicksberg's Theorem (Glicksberg, 1952))**.** *Let G be a continuous game, in the sense described above. Then G has a mixed strategy Nash equilibrium.*

[5]The players' mixed strategies are assumed to be independent.

Proof Idea. Though Glicksberg's own proof relies on a generalisation of the fixed-point theorem used to prove Nash's existence theorem, an alternative proof given in Ozdaglar (2010, Lecture 6) makes use of Nash's existence theorem without any further fixed-point result. In particular, the continuous game is approximated with a sequence of finite games, each corresponding to successively finer discretisations of the original game. Nash's existence theorem produces an equilibrium for each approximation, which we can show using the weak topology and the continuity assumptions to converge to an equilibrium of the original game. The full proof is given in Section B.1. □

4.3 Game-Theoretic Results for GANs

4.3.1 Mixing GANs

Two immediate problems arise with applying Glicksberg's Theorem to GANs. Firstly, we may for practical purposes be concerned about the *support size* of the mixed strategy. In other words: how many generators does it take to fool a discriminator of a certain strength? A classical result tells us that, so long as our generators are capable of producing simple Gaussian distributions, we can arbitrarily approximate p_r. Yet if the support size necessary to achieve this is too large, then training so many generators will be computationally prohibitive.

Secondly, it is not clear what it *means* for a generator to employ a mixed strategy. Of course, the generator induces a probability distribution p_G which it aims to be close to p_r. However, this is not the relevant probability distribution when considering mixed strategies in the GAN game. Instead, a mixed strategy would be a distribution over the *parameters* of G - in other words, a mixed strategy is a probability distribution over functions that induce probability distributions!

Arora et al. (2017) show how a mixture of finitely many generators and discriminators may approximate the minimax solution of the GAN game. Firstly, we define such an approximate equilibrium.

Definition 4.3.1. Let $\langle(\Delta(A_i), (U_i)\rangle_{i=1,2}$ be a mixed strategic two-player game, and let $\varepsilon > 0$. A mixed strategy profile $\alpha^* = (\alpha_1^*, \alpha_2^*)$ is an **ε-approximate equilibrium for** G if, for some $i = 1, 2$,

$$U_i(\alpha^*) \geq U_i(\alpha_i', \alpha_{-i}^*) - \varepsilon \text{ for all } \alpha_i' \in \Delta(A_i). \tag{4.3.1}$$

This leads Arora et al. to prove a theorem showing both the existence of an ε-approximate equilibrium using a finite mixture of generators and single discriminator, and a procedure for constructing an ε-approximate *pure* equilibrium. In doing so, we address the two concerns above. This theorem holds for the **neural net divergence**, defined here.

Definition 4.3.2 (Arora et al. (2017))**.** Let $\mathcal{X} \subseteq \mathbb{R}^d$ be an ambient data space, let p_G be the distribution induced by our generator function, and let p_r be the distribution of real data points over $\mathcal{X}$. Let $\mathcal{F}$ be a class of neural networks from $\mathbb{R}^d$ to $[0, 1]$, such that if $f \in \mathcal{F}$, then $1 - f \in \mathcal{F}$. Let ϕ be a measuring function: that is, $\phi\colon [0, 1] \to \mathbb{R}$ be concave and monotone. The **neural $\mathcal{F}$-divergence** with respect to ϕ between two distributions μ, ν supported on $\mathcal{X}$ is defined as

$$d_{\mathcal{F},\phi}(\mu\|\nu) := \sup_{D\in\mathcal{F}} \big(-J_\phi^D(D, \nu)\big), \tag{4.3.2}$$

where

$$J_\phi^D(D,\nu) = -\Big(\mathbb{E}_{x\sim\mu}[\phi(D(x))] + \mathbb{E}_{y\sim\nu}[\phi(1-D(y))]\Big) \tag{4.3.3}$$

Remark 4.3.3. Note that, in seeking a generator G that minimises

$$d_{\mathcal{F},\phi}(p_r\|p_G),$$

we are seeking a solution to a zero-sum game, with the loss functions defined as expected.

Furthermore, the neural net divergence may be seen as a generalisation of the practical instantiations of the JS and Wasserstein distance (setting $\phi(x) = \log(x)$ or $\phi(x) = x$, respectively), where the term 'practical' denotes the fact that our function space consists of neural networks. As such, any results we can prove about minimising neural net divergence will be highly relevant to the GAN examples most prominently considered so far.

We are now in a position to give the main positive result with regards to GANs and the existence of equilibria.

Theorem 4.3.4 (Arora et al. (2017), Theorem 4.3)**.** *Let ϕ be an L_ϕ-Lipschitz concave measuring function bounded in $[-\Delta,\Delta]$, and suppose the generator and discriminators are L-Lipschitz with respect to the parameters and L'-Lipschitz with respect to inputs. Furthermore, suppose the generator can approximate any point mass: that is, for all points x and any $\epsilon > 0$, there is a generator such that $\mathbb{E}_{z\sim p_z}[\|G(z)-x\|] \leq \epsilon$.*

Suppose the generator and discriminator are both k-layer neural networks ($k \geq 2$) with p parameters, and the last layer uses the ReLU activation function $f(x) = \max\{0,x\}$. Then there exists $(k+1)$-layer neural networks of generators G and discriminator D with $O\left(\frac{\Delta^2 p^2 \log(LL'L_\phi \cdot p/\varepsilon)}{\varepsilon^2}\right)$ parameters, such that there exists an ε-approximate pure equilibrium with value $2\phi(1/2)$ to the game induced by $d_{\mathcal{F},\phi}(p_r\|p_G)$.

Proof. The proof can be found in Appendix B, Section B.2. □

We have now seen how, when we consider GANs *qua* strategic zero-sum two-player games, we are able to establish the existence of approximate solutions to a broad variety of GAN games even when taking into account the limitations of the generator and discriminator functions.

Of course, non-zero-sum games such as the NS-GAN are not covered by this result, which only applies when the generator and discriminator are optimising for the same value function. However, as shown previously, Glicksberg's Theorem guarantees the existence of mixed strategy Nash equilibria for such a game, given the plausible assumption that the parameter spaces of the neural network are compact. It is an open problem whether the procedure for producing a single neural network that simulates a 'mixed strategy' can be extended to guarantee the existence of pure equilibria in non-zero-sum games, or whether we could get a realistic bound for the support size of such a mixture in exchange for settling for an ε-approximate equilibrium.

4.4 Frontiers of Research

In this final section, I will briefly review some of the ongoing research in game theory that is applicable to better understanding the dynamics and possible solutions of GAN games.

4.4.1 Existence does not Guarantee Convergence

As Arora et al. (2017, p. 15) observe, in the practical cases in which $V(\cdot, \cdot)$ is not convex-concave, the mere existence of an equilibrium does not guarantee that a simple algorithm like gradient descent will converge to it. The following is a pathological example.[6] for a non-GAN zero-sum game that fails to converge to its equilibrium using gradient descent.

Example 4.4.1 (Goodfellow (2016), Sections 7.2 and 8.2)**.** Consider a zero-sum game with two players that each control a single scalar value. The minimising player controls scalar x, the maximising player controls scalar y, and the value function for the game is

$$V(x, y) = xy. \tag{4.4.1}$$

By solving $\partial_x V(x, y) = 0$ and $\partial_y V(x, y) = 0$, we can establish that there is a saddle point at $x = y = 0$. Moreover, this saddle point is a Nash equilibrium: if the minimising player fixes $x = 0$, the maximising player cannot attain a better score than at $y = 0$, and vice versa.

Now, suppose the players were to learn this equilibrium via SimGD. To simplify the problem, we imagine that gradient descent is a continuous time process with an infinitesimal learning rate, so that the SimGD is described by the following system of partial differential equations:

$$\frac{\partial x}{\partial t} = -\frac{\partial}{\partial x} V(x(t), y(t)), \tag{4.4.2}$$

$$\frac{\partial y}{\partial t} = \frac{\partial}{\partial y} V(x(t), y(t)). \tag{4.4.3}$$

Clearly, these evaluate to

$$\frac{\partial x}{\partial t} = -y(t), \tag{4.4.4}$$

$$\frac{\partial y}{\partial t} = x(t). \tag{4.4.5}$$

Differentiating (4.4.5) yields

$$\frac{\partial^2 y}{\partial t^2} = \frac{\partial x}{\partial t} = -y(t). \tag{4.4.6}$$

This differential equation has the solution

$$x(t) = x(0)\cos(t) - y(0)\sin(t) \tag{4.4.7}$$

$$y(t) = x(0)\sin(t) + y(0)\cos(t). \tag{4.4.8}$$

[6]For further examples, refer to Arora et al. (2017, Appendix C) Moreover, Mertikopoulos et al. (2018) proves the existence of cycling behaviour for two players adopting another approach for finding the solution to a zero-sum game.

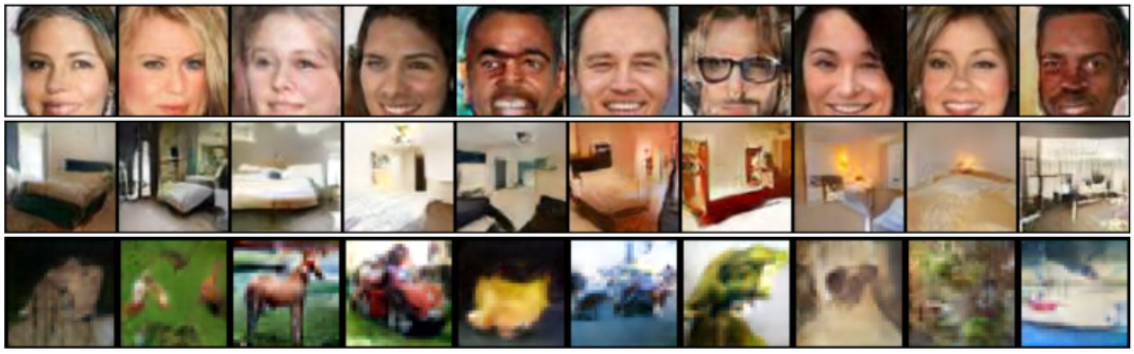

Figure 4.1: Images sampled from a Coulomb GAN after training on the CelebA (first row), LSUN bedrooms (second row), and CIFAR 10 (last row) datasets. Image taken from Unterthiner et al. (2017).

These dynamics form a circular orbit, so that SimGD with an infinitesimal learning rate will orbit around the equilibrium forever, at the same radius that it was initialised. Moreover, a larger learning rate can overshoot, causing SimGD to spiral outward forever. Hence, unless the players find themselves at the equilibrium *upon initialisation*, they will never approach the equilibrium using SimGD.

4.4.2 Convergence to Other Types of Equilibria

Some recent work has established the existence of other types of equilibria that a GAN will be guaranteed to converge to, or other types of GAN that are guaranteed to converge to an equilibrium.

As an example of the former, Hazan et al. (2017) defines a natural notion of *regret* for the players of a GAN game, and gives gradient-based methods that guarantee convergence to a newly-defined notion of local, regret-based equilibrium. Though the paper specifically refers to MM-GAN games, the result can in fact be generalised to any GAN game whose loss functions are bounded, Lipschitz, and have Lipschitz gradients.

It remains to be seen, however, whether a generator-discriminator pair that attains such an equilibrium produces samples that synthesise p_r well. Could there be a relation that holds between p_r and p_G when an equilibrium is reached? Recall, in the case of the IMM-GAN, the minimax solution coincided with the two distributions in fact being equal.

As an example of the latter, Unterthiner et al. (2017) introduces a new version of the GAN problem which treats the data samples as charged particles on a potential field. This GAN model possesses only one Nash equilibrium, which is optimal in that we get $p_r = p_G$.

However, this approach suffers from the drawback that it requires the discriminator to learn slow enough to accurately estimate the potential function induced by the generator, and that the generator must in turn learn even more slowly.[7] Moreover, the samples it produces for some standard datasets are not particularly visually appealing (see Figure 4.1).

[7] Unterthiner et al. (2017, p. 7).

Chapter 5

Conclusion

The literature on GANs is still in a very early stage, with the vast majority of papers developing their theory having been published within the last 18 months. As we have seen, there is still some way to go in understanding GAN training: what it means for GANs to perform well, and what guarantees there are on whether a GAN will perform well.

Nonetheless, both the topological and game-theoretic perspectives on GANs allow for us to propose versions of GANs that avoid the immediate pitfalls like mode collapse. Moreover, they allow us to make use of the theoretical insights from these areas: for instance, the Wasserstein GAN is inspired by previous research in optimal transport.

There is still plenty of work to be done on GANs: the game-theoretic approach seems particularly promising and under-explored. It seems, moreover, that any progress made with training GANs using this approach would require tools in algorithmic game theory that would have applications elsewhere in the field.

Appendix A

Omitted Proofs from Chapter 3

Proof of Theorem 3.2.9, Part 1.

($\Rightarrow$) Let μ_m be the mixture distribution $\mu_m = \frac{1}{2}(\mu_n + \mu)$, so μ_m depends on n. For any signed measure ν, define $\|\nu\|_{\mathrm{TV}} := \sup_{A \subseteq \mathcal{X}} |\nu(A)|$ for all Borel sets A. Then

$$\begin{aligned} d_{\mathrm{TV}}\left(\mu_m \| \mu_n\right) &= \|\mu_m - \mu_n\|_{\mathrm{TV}} \\ &= \frac{1}{2}\|\mu - \mu_n\|_{\mathrm{TV}} \\ &= \frac{1}{2} d_{\mathrm{TV}}\left(\mu_n \| \mu\right) \leq d_{\mathrm{TV}}\left(\mu_n \| \mu\right). \end{aligned}$$

Hence, $d_{\mathrm{TV}}\left(\mu_m \| \mu_n\right)$ tends to 0 if $d_{\mathrm{TV}}\left(\mu_n \| \mu\right)$ tends to 0.

Now let $f_n = \frac{\mathrm{d}\mu_n}{\mathrm{d}\mu_m}$ be the Radon-Nykodim derivative between μ_n and μ_m. By construction, we have for every Borel set A that $\mu_n(A) \leq 2\mu_m(A)$. Picking $A = \{f_n > 3\}$, we get

$$\mu_n(A) = \int_A f_n \, \mathrm{d}\mu_m \geq 3\mu_m(A),$$

and so by these two inequalities, $\mu_m(A) = 0$. Hence $f_n \leq 3$ almost everywhere, with respect to the measures μ_m, μ_n, and μ.

Now fix $\varepsilon > 0$, and $A_n = \{f_n > 1 + \varepsilon\}$. Then

$$\mu_n(A_n) = \int_{A_n} f_n \, \mathrm{d}\mu_m \geq (1 + \varepsilon)\mu_m(A_n).$$

Hence

$$\begin{aligned} \varepsilon\mu_m(A_n) &\leq \mu_n(A_n) - \mu_m(A_n) \\ &\leq |\mu_n(A_n) - \mu_m(A_n)| \\ &\leq d_{\mathrm{TV}}\left(\mu_n \| \mu_m\right) \\ &\leq d_{\mathrm{TV}}\left(\mu_n \| \mu\right). \end{aligned}$$

Furthermore,

$$\begin{aligned}\mu_n(A_n) &\leq \mu_m(A_n) + |\mu_n(A_n) - \mu_m(A_n)| \\ &\leq \frac{1}{\varepsilon} d_{\mathrm{TV}}\left(\mu_n \| \mu\right) + d_{\mathrm{TV}}\left(\mu_n \| \mu_m\right) \\ &\leq \left(\frac{1}{\varepsilon} + 1\right) d_{\mathrm{TV}}\left(\mu_n \| \mu\right).\end{aligned}$$

Therefore,

$$\begin{aligned}d_{\mathrm{KL}}\left(\mu_n \| \mu_m\right) &= \int_{\mathcal{X}} \log(f_n)\, \mathrm{d}\mu_n \\ &\leq \log(1+\varepsilon) + \int_{A_n} \log(f_n)\, \mathrm{d}\mu_n \\ &\leq \log(1+\varepsilon) + \log(3)\mu_n(A_n) \\ &\leq \log(1+\varepsilon) + \log(3)\left(\frac{1}{\varepsilon} + 1\right) d_{\mathrm{TV}}\left(\mu_n \| \mu\right).\end{aligned}$$

By taking lim sup on both sides, we get that

$$0 \leq \limsup d_{\mathrm{KL}}\left(\mu_n \| \mu_m\right) \leq \log(1+\varepsilon)$$

for all $\varepsilon > 0$, and so $d_{\mathrm{KL}}\left(\mu_n \| \mu_m\right) \to 0$.

Likewise, we define $g_n = \frac{\mathrm{d}\mu}{\mathrm{d}\mu_m}$ and $B_n = \{g_n > 1+\varepsilon\}$, showing *mutatis mutandis* that $d_{\mathrm{KL}}\left(\mu \| \mu_m\right) \to 0$. From this, we conclude that

$$d_{\mathrm{JS}}\left(\mu_n \| \mu\right) = \frac{1}{2}\big(d_{\mathrm{KL}}\left(\mu_n \| \mu_m\right) + d_{\mathrm{KL}}\left(\mu \| \mu_m\right)\big) \to 0.$$

($\Leftarrow$) Using the triangle and Pinsker's inequalities, we get

$$\begin{aligned}d_{\mathrm{TV}}\left(\mu_n \| \mu\right) &\leq d_{\mathrm{TV}}\left(\mu_n \| \mu_m\right) + d_{\mathrm{TV}}\left(\mu \| \mu_m\right) \\ &\leq \sqrt{\frac{1}{2} d_{\mathrm{KL}}\left(\mu_n \| \mu_m\right)} + \sqrt{\frac{1}{2} d_{\mathrm{KL}}\left(\mu \| \mu_m\right)} \\ &\leq 2\sqrt{\frac{1}{2} d_{\mathrm{JS}}\left(\mu_n \| \mu\right)} \to 0.\end{aligned}$$

□

Proof of Theorem 3.2.13. Define

$$\begin{aligned}V(\tilde{f}, \theta) &:= \mathbb{E}_{x \sim p_r}[\tilde{f}(x)] - \mathbb{E}_{x \sim g_\theta}[\tilde{f}(x)] \\ &= \mathbb{E}_{x \sim p_r}[\tilde{f}(x)] - \mathbb{E}_{z \sim p_z}[\tilde{f}(g_\theta(x))],\end{aligned}$$

where $\tilde{f} \in \mathcal{F} = \mathrm{Lip}_1(\mathcal{X})$ and $\theta \in \mathbb{R}^d$.

Since $\mathcal{X}$ is compact, we know by the Kantorovich-Rubinstein duality (Theorem 3.2.2) that there is an $f \in \mathcal{F}$ such that

$$d_{\mathrm{W}}\left(p_r \| g_\theta\right) = \sup_{\tilde{f} \in \mathcal{F}} V(\tilde{f}, \theta) = V(f, \theta). \tag{A.0.1}$$

Now define $X^*(\theta)$ to be the set of f such that (A.0.1) holds. The duality tells us that this set is non-empty. Moreover, by an envelope theorem (Milgrom and Segal, 2002, Theorem 1) we have

$$\nabla_\theta d_{\mathrm{W}}\left(p_r \| g_\theta\right) = \nabla_\theta V(f, \theta) \tag{A.0.2}$$

for any $f \in X^*(\theta)$, when both terms are well-defined.

Now take an $f \in X^*(\theta)$. Then, when the first two terms are well-defined,

$$\begin{aligned}\nabla_\theta d_{\mathrm{W}}\left(p_r \| g_\theta\right) &= \nabla_\theta V(f, \theta) \\ &= \nabla_\theta[\mathbb{E}_{x\sim p_r}[f(x)] - \mathbb{E}_{z\sim p_z}[f(g_\theta(x))]] \\ &= -\nabla_\theta \mathbb{E}_{z\sim p_z}[f(g_\theta(x))].\end{aligned}$$

The technical remainder of this proof is to show that

$$-\nabla_\theta \mathbb{E}_{z\sim p_z}[f(g_\theta(x))] = -\mathbb{E}_{z\sim p_z}[\nabla_\theta f(g_\theta(x))]. \tag{A.0.3}$$

Since $f \in \mathcal{F}$, f is 1-Lipschitz. Moreover, $f(g_\theta(z))$ is locally Lipschitz on (θ, z) with constants $L(\theta, z)$ given by the assumption on g. Hence, by Rademacher's Theorem (Federer, 2014, Theorem 3.1.6), $f(g_\theta(z))$ is differentiable almost everywhere for (θ, z) jointly. In other words, the set $A = \{(\theta, z) \mid f \circ g \text{ is not differentiable}\}$ has measure 0.

By Fubini's Theorem, this implies that for almost every θ, the section $A_\theta = \{z \mid (\theta, z) \in A\}$ has measure 0. Fix some θ_0 such that the measure of A_{θ_0} is null, and the RHS of equation (A.0.3) is well-defined. For this θ_0, we have that $\nabla_\theta f(g_\theta(z))|_{\theta_0}$ is well-defined for almost any z, and p_z-almost everywhere.

By our Lipschitz assumption, we know that

$$\mathbb{E}_{z\sim p_z}[\|\nabla_\theta f(g_\theta(z))|_{\theta_0}\|] \leq \mathbb{E}_{z\sim p_z}[L(\theta_0, z)] < +\infty, \tag{A.0.4}$$

so $\mathbb{E}_{z\sim p_z}[\nabla_\theta f(g_\theta(z))|_{\theta_0}]$ is well-defined for almost every θ_0. Therefore,

$$\begin{aligned}&\frac{\mathbb{E}_{z\sim p_z}[f(g_\theta(z))] - \mathbb{E}_{z\sim p_z}[f(g_{\theta_0}(z))] - \left\langle(\theta - \theta_0, \mathbb{E}_{z\sim p_z}[\nabla_\theta f(g_\theta(z))|_{\theta_0}]\right\rangle}{\|\theta - \theta_0\|} \\ &\qquad = \mathbb{E}_{z\sim p_z}\left[\frac{f(g_\theta(z)) - f(g_{\theta_0}(z)) - \left\langle(\theta - \theta_0, \nabla_\theta f(g_\theta(z))|_{\theta_0}\right\rangle}{\|\theta - \theta_0\|}\right].\end{aligned} \tag{A.0.5}$$

By the differentiability of $f \circ g$, the term inside the integral converges p_z-almost everywhere to 0 as $\theta \to \theta_0$. Moreover,

$$\begin{aligned}&\left\|\frac{f(g_\theta(z)) - f(g_{\theta_0}(z)) - \left\langle(\theta - \theta_0, \nabla_\theta f(g_\theta(z))|_{\theta_0}\right\rangle}{\|\theta - \theta_0\|}\right\| \\ &\qquad \leq \frac{\|\theta - \theta_0\| L(\theta_0, z) + \|\theta - \theta_0\| \cdot \|\nabla_\theta f(g_\theta(z))|_{\theta_0}\|}{\|\theta - \theta_0\|} \\ &\qquad \leq 2L(\theta_0, z).\end{aligned} \tag{A.0.6}$$

Furthermore, since $\mathbb{E}_{z\sim p_z}[2L(\theta_0, z)] < +\infty$ by the Lipschitz assumption, by dominated convergence equation A.0.5 tends to 0 as $\theta \to \theta_0$, so that

$$\nabla_\theta \mathbb{E}_{z\sim p_z}[f(g_\theta(x))] = \mathbb{E}_{z\sim p_z}[\nabla_\theta f(g_\theta(x))] \tag{A.0.7}$$

for almost every θ, and in particular when the RHS is well-defined. The LHS is also proven to exist simultaneously. □

Appendix B

Omitted Proofs from Chapter 4

B.1 Proof of Theorem 4.2.7 (Glicksberg's Theorem)

This proof is as in Ozdaglar (2010, Lecture 6). We first give a generalised definition of a continuous game.

Definition B.1.1. A **continuous game** is a game $\langle \mathcal{I}, (S_i), (u_i) \rangle$ where $\mathcal{I}$ is a finite set, the S_i are non-empty compact metric spaces, and the $u_i \colon S \to \mathbb{R}$ are continuous functions valued on $S = \times_{i \in \mathcal{I}} S_i$.

Let $u = (u_1, ..., u_I)$ and $\tilde{u} = (\tilde{u}_1, ..., \tilde{u}_I)$ be two profiles of utility functions ($|\mathcal{I}| = I$) defined on S such that, for each $i \in \mathcal{I}$, the functions $u_i, \tilde{u}_i$ are bounded and measurable. We can define the distance between the utility function profiles u and $\tilde{u}$ as

$$\max_{i \in \mathcal{I}} \sup_{s \in S} |u_i(s) - \tilde{u}_i(s)|. \tag{B.1.1}$$

Clearly, this distance is symmetric, obeys the triangle inequality, and is positive definite. Let $G = \langle \mathcal{I}, (S_i), (u_i) \rangle$ and $\tilde{G} = \langle \mathcal{I}, (S_i), (\tilde{u}_i) \rangle$ be two games, and let σ be an equilibrium of G. We can show that, if u and $\tilde{u}$ are close with respect to the distance given above, σ is an ε-equilibrium of $\tilde{G}$. Here, the definition of a ε-equilibrium for G generalises the definition given for a two-player game (Definition 4.3.1).

Lemma B.1.2. *Let G be a continuous game. Assume that $\sigma^k \to \sigma$ and $\varepsilon^k \to \varepsilon$ as $k \to \infty$, where for each k, σ^k is an ε^k-equilibrium of G. Then σ is an ε-equilibrium of G.*

Proof. We have by definition

$$u_i(s_i, \sigma^k_{-i}) \leq u_i(\sigma^k) + \varepsilon^k \quad \forall i \in \mathcal{I},\ \forall s_i \in S_i. \tag{B.1.2}$$

Taking the limit as $k \to \infty$ in (B.1.2), and using the continuity of the utility functions together with the convergence of the mixture distributions under the weak topology, we obtain

$$u_i(s_i, \sigma_{-i}) \leq u_i(\sigma) + \varepsilon \quad \forall i \in \mathcal{I},\ \forall s_i \in S_i. \tag{B.1.3}$$

Hence σ is an ε-equilibrium of G. □

The next step is to define a notion of closeness of two strategic games, given the distance of utility function.

Definition B.1.3. Let $G = \langle \mathcal{I}, (S_i), (u_i) \rangle$ and $G' = \langle \mathcal{I}, (S_i), (u_i') \rangle$ be two strategic games. We say that G' is an α-**approximation to** G if for all $i \in \mathcal{I}$ and $s \in S$, we have

$$|u_i(s) - u_i'(s)| \leq \alpha. \tag{B.1.4}$$

Lemma B.1.4. *If G' is an α-approximation to G and σ is an ε-equilibrium of G', then σ is an $(\varepsilon + 2\alpha)$-equilibrium of G.*

Proof. For all $i \in \mathcal{I}$ and $s_i \in S_i$, we have

$$\begin{aligned} u_i(s_i, \sigma_{-i}) - u_i(\sigma) &= \big(u_i(s_i, \sigma_{-i}) - u_i'(s_i, \sigma_{-i})\big) \\ &\quad + \big(u_i'(s_i, \sigma_{-i}) - u_i'(\sigma)\big) \\ &\quad + \big(u_i'(\sigma) - u_i(\sigma)\big) \\ &\leq \alpha + \varepsilon + \alpha = \varepsilon + 2\alpha. \end{aligned}$$

□

The next proposition gives us the 'discretisation' necessary to approximate our continuous game to an arbitrary degree of accuracy.

Lemma B.1.5. *For any continuous game G and any $\alpha > 0$, there exists an 'essentially finite' game which is an α-approximation to G.*

Proof. Since S is a compact metric space with metric d, the utility functions u_i are uniformly continuous. That is, for all $\alpha > 0$, there exists some $\varepsilon > 0$ such that whenever $d(s,t) \leq \varepsilon$,

$$|u_i(s) - u_i(t)| \leq \alpha. \tag{B.1.5}$$

Moreover, since S_i is compact, it can be covered with finitely many open balls U_i^j, each with radius less than ε. We assume without loss of generality that these balls are disjoint and non-empty.

Now pick some $s_i^j \in U_i^j$ for each i, j. Then we define our 'essentially finite' game G' with utility functions u_i' to be given by

$$u_i'(s) = u_i(s_1^j, ..., s_I^j), \quad \forall s \in U^j = \times_{k=1}^I U_k^j. \tag{B.1.6}$$

This game is 'essentially finite', in that while the utility functions are valued on all of S, they nonetheless attain only finitely many values.

Then, for all $s \in S$ and all $i \in \mathcal{I}$, we have by uniform continuity that

$$|u_i'(s) - u_i(s)| \leq \alpha, \tag{B.1.7}$$

since $d(s, s^j) \leq \varepsilon$ for all j. This gives us the desired result. □

We can now prove Glicksberg's Theorem.

Proof of Theorem 4.2.7. Let (α^k) be a sequence such that $\alpha^k \to 0$ as $k \to \infty$. By Lemma B.1.5, there exists for each α^k an α^k-approximation G^k of G.

Since each G^k is 'essentially finite' for each k, we can use Nash's existence theorem to guarantee the existence of a Nash equilibrium, equivalently a 0-approximate Nash equilibrium. Denote this by σ^k. Then, by Lemma B.1.4, σ^k is a $2\alpha^k$-equilibrium of G.

Since S is compact, the space of mixed distributions $\Delta(S)$ is compact, so $\{\sigma^k\}$ has a convergent subsequence. Without loss of generality, assume that $\sigma^k \to \sigma$. Since $2\alpha^k \to 0$ and $\sigma^k \to \sigma$ as $k \to \infty$, it follows by Lemma B.1.2 that σ is a 0-approximate equilibrium, hence a Nash equilibrium, of G. □

B.2 Proof of Theorem 4.3.4

We must first show that there is a finite mixture of generators and discriminators that can approximate the equilibrium that exists for infinite mixtures.

Suppose ϕ is an L_ϕ-Lipschitz concave measuring function bounded in $[-\Delta, \Delta]$. Let $\mathcal{U}$, $\mathcal{V} \subseteq \mathbb{R}^p$ be the (compact) parameter spaces of the generator and discriminator, respectively. Suppose the generator and discriminators are L-Lipschitz with respect to the parameters and L'-Lipschitz with respect to inputs. Suppose further the value function F for the minimax game is given by the neural $\mathcal{F}$-divergence[1] with respect to ϕ:

$$F(u, v) = \mathbb{E}_{x\sim p_r}[\phi(D_v(x))] + \mathbb{E}_{z\sim p_z}[\phi(1 - D_v(G_u(z))]. \tag{B.2.1}$$

Furthermore, we suppose the generator can approximate any point mass: that is, for all points x and any $\varepsilon > 0$, there is a generator such that $\mathbb{E}_{z\sim p_z}[\|G(z) - x\|] \leq \varepsilon$.

Note: all of the proofs within this section are as in Arora et al. (2017, Appendix B).

Lemma B.2.1 (Arora et al. (2017), Theorem 4.2)**.** *In the above setting, there exists a universal constant $C > 0$ such that for any $\varepsilon > 0$, there exists*

$$T = \frac{C\Delta^2 p \log(LL'L_\phi p/\varepsilon)}{\varepsilon^2}$$

generators $G_{u_1}, ..., G_{u_T}$ such that, if $\mathcal{S}_u$ is the uniform distribution on u_i, and D is a discriminator that outputs only 1/2, then $(\mathcal{S}_u, D)$ is an ε-approximate Nash equilibrium.

Proof. Let V be the value of the associated zero-sum game. One strategy of the discriminator is to just output 1/2. Since this strategy has payoff $2\phi(1/2)$ no matter what the generator does, it follows that $V \geq 2\phi(1/2)$.

By assumption, for any point x and any $\varepsilon > 0$, there is a generator $G_{x,\varepsilon}$ such that $\mathbb{E}_{z\sim p_z}[\|G_{x,\varepsilon}(z) - x\|] \leq \varepsilon$. For any $\zeta > 0$, we can take a sample $x \sim p_r$, and use the generator $G_{x,\zeta}$. Let p_ζ be the distribution generated by this mixture of generators. By the point mass approximation property of the generators, $d_{\mathrm{W}}(p_r \| p_\zeta) \leq \zeta$. Combining this with the fact that the discriminator is L'-Lipschitz and that ϕ is L_ϕ-Lipschitz, it holds for any discriminator D_v that

$$|\mathbb{E}_{x\sim p_r}[\phi(1 - D(x))] - \mathbb{E}_{x\sim p_\zeta}[\phi(1 - D(x))]| \leq O(L_\phi L'\zeta). \tag{B.2.2}$$

Therefore,

$$\begin{aligned}
&\max_{v\in\mathcal{V}}\Big(\mathbb{E}_{x\sim p_r}[\phi(D(x))] + \mathbb{E}_{x\sim p_\zeta}[\phi(1 - D(x))]\Big) \\
&\leq \max_{v\in\mathcal{V}}\Big(\mathbb{E}_{x\sim p_r}[\phi(D(x)) + \phi(1 - D(x))]\Big) + O(L_\phi L'\zeta) \\
&\leq 2\phi(1/2) + O(L_\phi L'\zeta).
\end{aligned}$$

Here, the last step uses the assumption that ϕ is concave. Hence $V \leq 2\phi(1/2) + O(L_\phi L'\zeta)$ for any ζ. Letting $\zeta \to 0$, we get that $V = 2\phi(1/2)$.

[1] Here, $\mathcal{F} = \{D_v \mid v \in \mathcal{V}\}$.

Now let $(\mathcal{S}'_u, \mathcal{S}'_v)$ be the pair of optimal mixed strategies as given by Glicksberg's Theorem (Theorem 4.2.7), and let V be the optimal value. We will show that, by randomly sampling T generators from $\mathcal{S}'_u$, we get the desired mixture with high probability.

First, we construct $(\frac{\varepsilon}{4LL'L_\phi})$-nets for the discriminator parameter space $\mathcal{V}$. Let A be the set of centres for such nets. Since $\mathcal{V}$ is compact, A is finite. Moreover, by a standard construction of such nets, we have that

$$\log(|A|) \leq C'n \log(LL'L_\phi \cdot p/\varepsilon)$$

for some constant C'. Let $u_1, ..., u_T$ be independent samples from $\mathcal{S}'_u$. By the Chernoff bound, for any net centre $a \in A$,

$$\mathbb{P}\left(\mathbb{E}_{i\in[T]}[F(u_i, a)] \geq \mathbb{E}_{u\in\mathcal{U}}[F(u, a)] + \varepsilon/2\right) \leq \exp\left(-\frac{\varepsilon^2 T}{2\Delta^2}\right). \tag{B.2.3}$$

Therefore, when $T = \frac{C\Delta^2 p \log(LL'L_\phi p/\varepsilon)}{\varepsilon^2}$ and the constant C is greater than $2C'$, then with high probability this inequality is true for all $a \in A$. For any $v \in \mathcal{V}$, let a' be the closest point in the net, so by construction $\|v - a'\| \leq \frac{\varepsilon}{4LL'L_\phi}$. Using this, one can derive straightforwardly that $F(u, v)$ is $(2LL'L_\phi)$-Lipschitz in both u and v, so that

$$\mathbb{E}_{i\in[T]}[F(u_i, a')] \leq \mathbb{E}_{i\in[T]}[F(u_i, v)] + \varepsilon/2. \tag{B.2.4}$$

We then get for any $v' \in \mathcal{V}$ that

$$\mathbb{E}_{i\in[T]}[F(u_i, v')] \leq 2\phi(1/2) + \varepsilon. \tag{B.2.5}$$

Therefore, this mixture of generators can win against any discriminator, and by a probabilistic argument, there must exist such generators. Since the discriminator given by constant $1/2$ can achieve the value V regardless of the generator value, we get an approximate equilibrium. □

Given the existence of the approximate equilibrium, the next step to prove the existence of an approximate *pure* equilibrium is to construct a single generator that can approximately generate the mixture distribution of generators. To do so, we pass our noise input $h \sim p_z$ through all the generators $G_{u_1}, ..., G_{u_T}$ and implement a 'multi-way selector' to select a uniformly random output from $G_{u_i}(h)$, where $i \in [T]$.

First, we show how to compute a step function using a two-layer neural network.

Lemma B.2.2 (Arora et al. (2017), Lemma 3)**.** *Let q be a positive integer and $z_1 < ... < z_q$. For any $0 < \delta < \min_{i\in[q-1]}(z_{i+1} - z_i)$, there is a two-layer neural network with single input $h \in \mathbb{R}$ that outputs $q + 1$ numbers $x_1, ..., x_{q+1}$ such that*

1. *$\sum_{i=1}^{q+1} x_i = 1$ for all h;*
2. *when $h \in [z_{i-1} + \delta/2, z_i - \delta/2]$, we have that $x_i = 1$ and all the other x_j are 0. When $h \leq z_1 - \delta/2$ only x_1 is 1, and when $h \geq z_q + \delta/2$ only $x_{q+1} = 1$.*

Proof. Using a two-layer neural network with ReLU activation functions, we can compute the function

$$f_i(h) = \max\left\{\frac{h - z_i - \delta/2}{\delta}, 0\right\} - \max\left\{\frac{h - z_i + \delta/2}{\delta}, 0\right\}. \tag{B.2.6}$$

This function evaluates to 0 for all $h < z_i - \delta/2$, and 1 for all $h \geq z_i + \delta/2$, and changes linearly in-between. Writing

$$\begin{aligned} x_1 &= 1 - f_1(h) \\ x_{q+1} &= f_q(h) \\ x_q &= f_i(h) - f_{i-1}(h) \quad \forall i \in \{2, 3, ..., q\}, \end{aligned}$$

we see that these functions satisfy (1)-(2). □

Using these step functions, we can design the multi-way selector.

Lemma B.2.3 (Arora et al. (2017), Lemma 4)**.** *In the setting above, suppose the generator and discriminator are both k-layer neural networks (where $k \geq 2$), and the last layer uses the ReLU activation function. Then there is a $(k+1)$-layer neural network with $O\left(\frac{\Delta^2 p^2 \log(LL' L_\phi \cdot p/\varepsilon)}{\varepsilon^2}\right)$ parameters that can generate a distribution within δ TV distance of the mixture of $G_{u_1}, ..., G_{u_T}$.*

Proof Idea. Since we have implemented step functions from Lemma B.2.2, we can pass the input through all the generators $G_{u_1}, ..., G_{u_T}$. At the last layer of each G_{u_i}, we add a large multiple of $-(1 - x_i)$, where x_i is the i-th outputof the network in Lemma B.2.2. Then, if $x_i = 0$, this will effectively 'de-activate' the network by bringing it below the threshold of the ReLU function. However, if $x_i = 1$, this will have no effect. By Lemma B.2.2, we know that most of the time only one of the x_i's will be 1, so that only one generator is selected. □

Proof. Suppose the input for the generator is $(h_0, h) \sim \mathcal{N}(0, 1) \times p_z$, where samples are drawn independently. We will pass the input h through the generators and get outputs $G_{u_i}(h)$, and then use h_0 to select one of these outputs as the 'true' output.

Let $z_1, ..., z_{T-1}$ be real numbers that divide the probability density of a Gaussian into T equal parts. Choose $\delta' = \delta/(100T)$ in Lemma B.2.2 to get a 2-layer neural net that computes step functions $x_1, ..., x_T$. Then the probability that $(x_1, ..., x_T)$ has more than one non-zero entry is smaller than δ. Now, for the output of $G_{u_i}(h)$, in each output ReLU gate, add a multiple of $-(1 - x_i)$ that is larger than than the maximum possible output. Therefore, when $x_i = 0$, the result before the ReLU will be negative and so the output will be 'disabled', and when $x_i = 1$ the output will be preserved. Call this modified network $\hat{G}_{u_i}$. Then $\hat{G}_{u_i} = G_{u_i}$ when $x_i = 1$ and $\hat{G}_{u_i} = 0$ when $x_i = 0$. Now add a layer that outputs the sum of $\hat{G}_{u_i}$.

By construction, when $(x_1, ..., x_T)$ has only one non-zero entry, the network correctly outputs the corresponding $G_{u_i}(x_i)$. The probability that this happens is at least $1 - \delta$, and so the TV distance with the mixture is bounded by δ. □

Theorem B.2.4 (Theorem 4.3.4 restated)**.** *In the setting above, there exists $(k+1)$-layer neural networks of generators G and discriminator D with $O\left(\frac{\Delta^2 p^2 \log(LL' L_\phi \cdot p/\varepsilon)}{\varepsilon^2}\right)$ parameters, such that there exists an ε-approximate pure equilibrium with value $2\phi(1/2)$ to the game induced by $d_{\mathcal{F},\phi}(p_r \| p_G)$.*

Proof. Let T be large enough so that there exists an $\varepsilon/2$-approximate mixed Nash equilibrium. Let the new set of generators be constructed as in Lemma B.2.3, with $\delta \leq \varepsilon/(4\Delta)$ and $G_{u_1}, ..., G_{u_T}$ as the original set of generators. Let D be the discriminator that always outputs $1/2$, and G be the 'multi-way selector' generator

constructed by the T generators from the approximate mixed equilibrium. Let $F^*(G, D)$ denote the value for the new two-player game.[2] For any discriminator D_v, we have

$$\begin{aligned} F^*(G, D_v) &\leq \mathbb{E}_{i \in [T]}[F(u_i, v)] + |F^*(G, D) - \mathbb{E}_{i \in [T]} F(u_i, v)| \\ &\leq V + \varepsilon/2 + 2\Delta \frac{\varepsilon}{4\Delta} \\ &\leq V + \varepsilon. \end{aligned}$$

Here, the bound for the first term comes from Lemma B.2.1, and the fact that the expectation is smaller than the maximum of expected values. The bound for the second term comes from the fact that changing a δ amount of probability mass can change the payoff F by at most $2\Delta\delta$ (recalling that ϕ is bounded in $[-\Delta, \Delta]$). Therefore, the generator will still fool all discriminators, and we therefore get the required pure equilibrium. □

[2]The game is new, since the space $\mathcal{F}$ of neural nets now includes neural nets of $(k+1)$ layers instead of just k layers.

Bibliography

M. Arjovsky and L. Bottou. Towards Principled Methods for Training Generative Adversarial Networks. *arXiv preprint arXiv:1701.04862*, 2017.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

S. Arora, R. Ge, Y. Liang, T. Ma, and Y. Zhang. Generalization and Equilibrium in Generative Adversarial Nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.

G. Basso. A Hitchhiker's Guide to Wasserstein Distances, 2015. URL `http://n.ethz.ch/~gbasso/download/A%20Hitchhikers%20guide%20to%20Wasserstein/A%20Hitchhikers%20guide%20to%20Wasserstein.pdf`.

O. Bousquet, S. Gelly, I. Tolstikhin, C.-J. Simon-Gabriel, and B. Schoelkopf. From Optimal Transport to Generative Modeling: the VEGAN Cookbook. *arXiv preprint arXiv:1705.07642*, 2017.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

D. A. Edwards. On the Kantorovich-Rubinstein Theorem. *Expositiones Mathematicae*, 29(4):387–398, 2011.

K. Fan. Minimax Theorems. *Proceedings of the National Academy of Sciences*, 39 (1):42–47, 1953.

H. Federer. *Geometric Measure Theory*. Springer, 2014.

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

I. L. Glicksberg. A Further Generalization of the Kakutani Fixed Point Theorem, with Application to Nash Equilibrium Points. *Proceedings of the American Mathematical Society*, 3(1):170–174, 1952.

I. Goodfellow. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv preprint arXiv:1701.00160*, 2016.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

E. Hazan, K. Singh, and C. Zhang. Efficient Regret Minimization in Non-Convex Games. *arXiv preprint arXiv:1708.00075*, 2017.

S. Kakutani. Concrete Representation of Abstract (m)-spaces (a Characterization of the Space of Continuous Functions). *Annals of Mathematics*, pages 994–1024, 1941.

T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive Growing of GANS for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*, 2017.

E. Kreyszig. *Introductory Functional Analysis with Applications*, volume 1. Wiley New York, 1978.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

N. Lei, K. Su, L. Cui, S.-T. Yau, and D. X. Gu. A Geometric View of Optimal Transportation and Generative Model. *arXiv preprint arXiv:1710.05488*, 2017.

J. Li, A. Madry, J. Peebles, and L. Schmidt. Towards Understanding the Dynamics of Generative Adversarial Networks. *arXiv preprint arXiv:1706.09884*, 2017.

S. Liu, O. Bousquet, and K. Chaudhuri. Approximation and Convergence Properties of Generative Adversarial Learning. In *Advances in Neural Information Processing Systems*, pages 5551–5559, 2017.

M. Lucic, K. Kurach, M. Michalski, S. Gelly, and O. Bousquet. Are GANs Created Equal? a Large-Scale Study. *arXiv preprint arXiv:1711.10337*, 2017.

P. Mertikopoulos, C. Papadimitriou, and G. Piliouras. Cycles in Adversarial Regularized Learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

L. Mescheder, S. Nowozin, and A. Geiger. The Numerics of GANs. *arXiv preprint arXiv:1705.10461*, 2017.

L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled Generative Adversarial Networks. *arXiv preprint arXiv:1611.02163*, 2016.

P. Milgrom and I. Segal. Envelope Theorems for Arbitrary Choice Sets. *Econometrica*, 70(2):583–601, 2002.

A. Müller. Integral Probability Metrics and their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

J. F. Nash. Equilibrium Points in n-person Games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

M. J. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

A. Ozdaglar. 6.254 Game Theory with Engineering Applications, Spring 2010. Massachusetts Institute of Technology: MIT OpenCourseWare, https://ocw.mit.edu. License: Creative Commons BY-NC-SA.

H. Ramsauer, M. Heusel, S. Hochreiter, B. Nessler, and T. Unterthiner. Two Time-Scale Update Rule for Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, pages 6608–6619, 2017.

T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016.

T. Unterthiner, B. Nessler, G. Klambauer, M. Heusel, H. Ramsauer, and S. Hochreiter. Coulomb GANs: Provably optimal Nash equilibria via potential fields. *arXiv preprint arXiv:1708.08819*, 2017.

C. Villani. *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

P. Zhang, Q. Liu, D. Zhou, T. Xu, and X. He. On the Discrimination-Generalization Tradeoff in GANs. *arXiv preprint arXiv:1711.02771*, 2017.