

# Understanding GAN Theory

Hadi Elzayn  
Applied Math and Computational Science  
University of Pennsylvania  
hads@sas.upenn.edu

September 28, 2018

Introduction

GANs as Games

Understanding GANs in the LQG setting

Generalization and diversity

A deeper dive into generalizability and discriminatory capacity

Introduction

GANs as Games<sup>1</sup>

Understanding GANS in the LQG setting<sup>2</sup>

Generalization and diversity<sup>3</sup>

A deeper dive into generalizability and discriminatory capacity<sup>4</sup>

---

<sup>1</sup>This section is mostly based on [2]

<sup>2</sup>This section is mostly based on [6]

<sup>3</sup>This section is mostly based on [2] and [9]

<sup>4</sup>This section is mostly based on [2] and [9]

# Introduction

- ▶ Do GANs recover results that we can achieve by other methods, or are they fundamentally different?
- ▶ When do GANs converge in polynomially many samples, and does an optimal discriminator always exist?
- ▶ Will GANs learn the entire distribution, or just a subset of it?
- ▶ When GANs do converge, will the generated distribution be close to the true distribution?
- ▶ What generalization guarantees can we make about GANs?

- ▶ Recall that we use *Generative Adversarial Nets* to learn a probability distribution over samples from a complex distribution
- ▶ In the original formulation, the Discriminator (D) and the Generator (G) play the following game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

- ▶ For fixed  $G$ , the optimal discriminator is

$$D_G^*(x) = \frac{p_{data}(x)}{p_{data}(x) + p_g(x)}$$

- ▶ Total Variation (TV) distance:

$$\delta(\mathbb{P}_r, \mathbb{P}_g) = \sup_{A \in \Sigma} |\mathbb{P}_r(A) - \mathbb{P}_g(A)|$$

- ▶ Kullback-Leibler (KL) divergence

$$KL(\mathbb{P}_r \parallel \mathbb{P}_g) = \int \log \left( \frac{P_r(x)}{P_g(x)} \right) P_r(x) d\mu(x)$$

- ▶ Jensen-Shannon (JS) divergence

$$JS(\mathbb{P}_r, \mathbb{P}_g) = KL(\mathbb{P}_r \parallel \mathbb{P}_m) + KL(\mathbb{P}_g \parallel \mathbb{P}_m)$$

where  $\mathbb{P}_m$  is  $(\mathbb{P}_r + \mathbb{P}_g)/2$ .

- ▶ Earth-Mover (EM) distance aka Wasserstein-1 distance:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} \|x - y\|$$

## GANs as Games



- ▶ View the generator as choosing a parameter  $u \in \mathcal{U}$  and discriminator as choosing a parameter  $v \in \mathcal{V}$
- ▶ In the original formulation, we define the payoff for the discriminator as

$$J^{(D)} = -\frac{1}{2} \mathbb{E}_{x \sim p_{data}} \log D(x) - \frac{1}{2} \mathbb{E}_z \log(1 - D(G(x)))$$

- ▶ We define the generator payoff as

$$J^{(G)} = -J^{(D)}$$

- ▶ In other words, this is a *zero-sum* game
- ▶ We can write the value of zero-sum games concisely: define the (scaled) payoff function as:

$$F(u, v) = \mathbb{E}_{x \sim \mathcal{D}_{real}} [\phi(D_v(x))] + \mathbb{E}_{x \sim \mathcal{D}_G} [\phi(1 - D_v(x))]$$

- ▶ The generator wishes to maximize  $F$ , while the discriminator wishes to minimize it

**Yes.**

Theorem (Von Neumann, 1928)

*Every finite zero-sum game has at least one mixed strategy equilibrium. That is, there exist a number  $V$  and a pair of distributions over actions  $(S_u, S_v)$  such that*

$$\forall v, \mathbb{E}_{u \sim S_u} [F(u, v)] \leq V$$

*and*

$$\forall u, \mathbb{E}_{v \sim S_v} [F(u, v)] \geq V$$

*We refer to  $V$  as the value of the game.*

- ▶ An  $\epsilon$ -approximate equilibrium is one in which players can't benefit by more than  $\epsilon$  by changing their strategies
- ▶ In this setting,  $\mathcal{S}_u, \mathcal{S}_v$  form an  $\epsilon$ -approximate equilibrium if for some value  $V$ :

$$\forall v \in \mathcal{V}, \mathbb{E}_{u \sim \mathcal{S}_u}[F(u, v)] \leq V + \epsilon$$

$$\forall u \in \mathcal{U}, \mathbb{E}_{v \sim \mathcal{S}_v}[F(u, v)] \geq V - \epsilon$$

- ▶ Notice that in practice, we stop GANs at an  $\epsilon$ -approximate equilibrium, so this is really the relevant case

- ▶ Intuition: if a deep net can approximate a Gaussian, and we could take infinite deep nets, then we could approximate any density<sup>5</sup> with an infinite number of generators
- ▶ Von Neumann's theorem guarantees only that a *mixed* strategy equilibrium exists, but folklore<sup>6</sup> says that any *infinite* mixture can be approximated with an appropriate *finite* mixture
- ▶ It turns out that a *finite* number of generators will suffice

---

<sup>5</sup>Since a classical result, e.g. [4] or [8] says that arbitrary probability distributions can be approximated by an infinite mixture of Gaussians

<sup>6</sup>Arora et. al. cite Lipton and Young, 1994 for a formal version of this claim, but technically their result applies to large games with a payoff matrix - that is, for mixtures over potentially exponentially many but still finite number of pure strategies. It turns out there is indeed folklore for the countable or continuous compact case.

In the same setting, we have the the following existence theorem:

## Theorem

*If the generator can approximate any point mass, there is a constant  $C > 0$  such that for any  $\epsilon$ , there exist  $T$  generators  $G_{u_1}, \dots, G_{u_T}$  (with  $T = C\Delta^2 p \log(LL'L_\phi p/\epsilon)/\epsilon^2$ ) such that if  $S_u$  is a uniform distribution on  $u_i$ , and  $D$  is a discriminator that outputs just  $\frac{1}{2}$ , then  $(S_u, D)$  is an  $\epsilon$ -approximate equilibrium.*

NB - this theorem just says that there exists a finite mixture of generator networks that can achieve the pure strategy approximate equilibrium in which the generator wins.

The proof is by an  $\epsilon$ -net argument and Chernoff bounds applied to sampling  $T$  generators from the mixed strategy.

But, there actually exists a *pure* approximate equilibrium too if we quadratically blow up the parameter size:

## Theorem

*Suppose the generator and discriminator are both  $k$ -layer neural networks ( $k \geq 2$ ) with  $p$  parameters, and the last layer uses ReLU activation functions. Then there exists  $k + 1$ -layer neural networks for generators  $G$  and discriminator  $D$  with  $\mathcal{O}\left(\frac{\Delta^2 p^2 \log(LL' L_\phi p / \epsilon)}{\epsilon^2}\right)$  parameters, such that there exists an  $\epsilon$ -approximate pure equilibrium with value  $2\phi(1/2)$ .*

The proof of this is just construction - turns out you can build a neural network that functions as a 'multi-way' selector to choose randomly among each of your generators, at the cost of a blowup in your parameter space.

## Understanding GANs in the LQG setting

- ▶ In order to better understand what GANs are actually doing, we consider a simple setting that is easy to analyze<sup>7</sup>.
- ▶ Our restrictions will be quite strong - in particular, the generator is *linear*, which is much simpler than practical settings
- ▶ We can think of this as maintaining the adversarial and generative in 'GANs' but for now dropping the 'nets' (or, if you like, requiring linear activation functions)
- ▶ Despite these simplifications, there are still generalization issues, and moreover, we want to understand what the framework is doing

---

<sup>7</sup>Full disclosure: the reviewers of this paper at ICLR, according to open review, had concerns that this setting is so constrained as to make analysis here non-representative. That may be true, but GANs are fairly hard to study in general so it seems like even a limited setting is progress.



- ▶ Recall that PCA is an unsupervised learning algorithm that finds a linear (possibly in non-linear kernel) mapping to a lower dimensional space
- ▶ The 'principle components' are the basis vectors for this subspace
- ▶ Can be viewed as selecting the the subspace with maximum variation or the subspace with which we can achieve minimum reconstruction error
- ▶ Turns out you can find the PCs by doing an eigenvector decomposition of the empirical covariance matrix or just getting the singular value decomposition of the observations

---

<sup>8</sup>Recapitulated from notes of Prof. Shivani Agarwal

- ▶ Data  $Y$  is generated by a  $d$ -dimensional multivariate Gaussian  $\mathcal{N}(\mu, \mathbf{K})$
- ▶ Our generator,  $G$ , will take Gaussian noise  $X$  of rank  $k \leq d$
- ▶ The generators will all be linear, that is they map input noise to sample output via *linear*  $g$
- ▶ Our loss will be Wasserstein-2 - that is, under the distance interpretation (see below) the goal will be to find  $g^*$ :

$$g^* \in \inf_{g(\cdot) \in \mathcal{G}} W_2^2(\mathbb{P}_Y, \mathbb{P}_g(X))$$

## Theorem

*Let  $Y \sim \mathcal{N}(0, K_Y)$  with  $K_Y$  full-rank. Let  $X \sim \mathcal{N}(0, I_k)$  with  $k \leq d$ . The optimal population GAN solution to the Wasserstein-2 minimization under linear  $\mathcal{G}$  is the  $k$ -PCA solution of  $Y$ .*

Proof intuition: If we want to find

$$Y^* := \inf_{\mathbb{P}_{\hat{Y}}, \text{supp } \hat{Y} \subseteq \mathcal{S}} W_2^2(\mathbb{P}_Y, \mathbb{P}_{\hat{Y}})$$

where  $\mathcal{S}$  is a subspace of  $\mathbb{R}^d$ , we just take the projection of  $Y$  to  $\mathcal{S}$ . If  $Y$  is multivariate Gaussian, then its projection will also be multivariate Gaussian in a lower dimension, and the subspace which gives a projection that is closest in distance is the one spanned by the principle components.

If we replace  $\mathbb{P}_Y$  with the empirical distribution of  $n$  samples drawn from  $Y$ ,  $\mathbb{Q}_Y^n$ , we get the *empirical GAN*.

## Theorem

*Under the same setup, the empirical constrained quadratic GAN optimization is also equivalent to empirical PCA.*

- ▶ There is a sharp difference in convergence guarantees depending on whether the generator and discriminator are constrained or unconstrained.
- ▶ If  $\mathcal{G}$  were unconstrained (i.e. didn't have to be linear), then  $W_2^2(\mathbb{P}_Y, \mathbb{Q}_Y^n) \rightarrow 0$  with high probability at the rate of  $\mathcal{O}(n^{-2/d}) = \mathcal{O}(2^{-2 \log(n)/d})$  i.e. getting small distance requires exponentially (in  $d$ ) many samples<sup>9</sup>
- ▶ On the other hand if  $\mathcal{G}$  is constrained to be linear, forces the generator not to memorize the training set
- ▶ The optimization becomes  $\min_{\mu, \mathbf{K}} W_2^2(\mathbb{Q}_Y^n, \mathcal{N}(\mu, \mathbf{K}))$ , which has optimal solution  $(0, \mathbf{K})$ , and  $\mu \rightarrow 0$  with  $\mathcal{O}(d/n)$ <sup>10</sup>.
- ▶ Still, the convergence of the of the optimal  $\mathbf{K}_n \rightarrow \mathbf{K}$  is slow, though - still at  $\mathcal{O}(n^{-2/d})$ .<sup>11</sup>

---

<sup>9</sup>Intuition: if the generator is unconstrained, it can memorize the empirical distribution - we'll see in more detail why this gives a bad bound

<sup>10</sup>This rate follows from interpreting Wasserstein/optimal transport in light of Rate Distortion theory.

- ▶ But now suppose we constraint the discriminator to be at most quadratic:

$$\psi(y) = y^\top \mathbf{A} y / 2$$

with  $\mathbf{A}$  positive semidefinite.

- ▶ This is without loss of generality, and the solution to the generator's problem remains the same - perform empirical PCA.
- ▶ But now, the payoff function can be rewritten in terms of the distance between two normal distributions (the true and empirical Gaussians of the data)
- ▶ This rate of convergence is just how fast the empirical convergence converges to the population covariance, which occurs at a rate of  $\tilde{O}(\sqrt{d/n})$
- ▶ While this setting is artificial, these results hint at a general phenomenon: there are tradeoffs to capacities of generators and discriminators. We will see more below

## Generalization and diversity

- ▶ Another interpretation of GANs is as minimizing the distance between the true distribution and that of the generator in some appropriate metric
- ▶ In the original GAN formulation, we can rewrite the game and substitute in the optimal discriminator to write the cost function of the generator as

$$\begin{aligned}C(G) &= \max_D V(G, D) \\&= \mathbb{E}_{x \sim p_{data}} \log \left[ \frac{p_{data}(x)}{p_{data}(x) + p_g(x)} \right] + \mathbb{E}_{x \sim p_g} \left[ \log \frac{p_g(x)}{p_{data}(x) + p_g(x)} \right] \\&= -\log 4 + KL \left( p_{data} \left\| \frac{p_{data} + p_g}{2} \right\| \right) + KL \left( p_g \left\| \frac{p_{data} + p_g}{2} \right\| \right) \\&= -\log 4 + 2JS(p_{data} \| p_g)\end{aligned}$$

- ▶ In other words, the goal of the Generator is to *minimize the Jensen-Shannon divergence* between the generated distribution and the true distribution



- ▶ More generally, for a class of functions  $\mathcal{F}$ , associate the quantity

$$d_{\mathcal{F}}(\mathbb{P}_r, \mathbb{P}_{\theta}) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{x \sim \mathbb{P}_r}[f(x)] - \mathbb{E}_{x \sim \mathbb{P}_{\theta}}[f(x)]|$$

Note that  $d_{\mathcal{F}}$  is an integral probability (pseudo)metric

- ▶ The *Wasserstein* distance  $d_W$  corresponds<sup>12</sup> to  $d_{\mathcal{F}}$  with  $\mathcal{F}$  the set of 1-Lipschitz functions.
- ▶ Implementing the WGAN procedure means minimizing this  $d_{\mathcal{F}}$
- ▶ More generally, various flavors of GAN frameworks turn out to correspond to varying choices of which integral probability metric to optimize

---

<sup>12</sup>This is not obvious - it follows from the Kantorovich-Rubinstein duality theorem. For a more accessible explanation, see [here](#).

- ▶ Recall that Arjovsky et. al. designed the WGAN in part to deal with the empirical phenomenon of *mode collapse* (duplicates or near-duplicates)
- ▶ This problem is a symptom of a broader cause: the generator can *win* the game without being close in a metric that we care about
- ▶ This can mean generating a distribution with much smaller support than the true distribution (meaning the discriminator could win the game if it had access to the *true* distribution)
- ▶ It turns out that using the Wasserstein distance (by itself) does not fix this worry

Call  $\mathcal{D}_{real}$  the true data distribution,  $\mathcal{D}_G$  the generated distribution,  $\hat{\mathcal{D}}_{real}$  to be a (uniform) distribution over some empirical sample drawn from  $\mathcal{D}_{real}$ , and  $\hat{\mathcal{D}}_G$  is a (uniform) distribution over a set of examples from the generated distribution<sup>13</sup>.

## Definition

We say that a generated distribution *generalizes* under the distance  $d(\cdot, \cdot)$  with generalization error  $\epsilon$  if (with high probability)

$$\left| d(\mathcal{D}_{real}, \mathcal{D}_G) - d(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_G) \right| \leq \epsilon$$

Notice that this is a statement about the difference between (empirical and real) distances between the generated and real distributions.

---

<sup>13</sup>Why should these be the quantities we care about? We don't have access to  $\mathcal{D}_{real}$ , and GANs don't give a parametric form for  $\mathcal{D}_G$ , so all we have are empirical estimates  $\hat{\mathcal{D}}_{real}$  and  $\hat{\mathcal{D}}_G$ . The hope is that optimizing on  $\hat{\mathcal{D}}_{real}$  and  $\hat{\mathcal{D}}_G$  will say something about  $\mathcal{D}_{real}$  and  $\mathcal{D}_G$ .

## Lemma

Let  $\mu$  be a uniform Gaussian distribution  $\mathcal{N}(0, \frac{1}{d}I)$  and  $\hat{\mu}$  be an empirical version of  $\mu$  with  $m$  examples. Then  $d_{JS}(\mu, \hat{\mu}) = \log 2$  and  $d_W(\mu, \hat{\mu}) \geq 1.1$ .

- ▶ Follow from fact that  $\mu$  is continuous while  $\hat{\mu}$  is discrete
- ▶ Implication 1: suppose  $\mathcal{D}_{real} = \mathcal{D}_G = \mu$ . Then  $d_{JS}(\mathcal{D}_{real}, \mathcal{D}_G) = 0$ . But  $d_{JS}(\hat{\mathcal{D}}_{real}, \hat{\mathcal{D}}_G) > 1$ . To see this note that with probability 1 the two empirical distributions have disjoint support.
- ▶ Implication 2: Suppose  $\mathcal{D}_{real} = \mu$  and  $\mathcal{D}_G = \hat{\mathcal{D}}_{real} = \hat{\mu}$ . (i.e.  $\mathcal{D}_G$  memorizes the training examples in  $\mathcal{D}_{real}$ . With enough (but polynomial) examples, we can make  $\hat{\mathcal{D}}_G \approx \mathcal{D}_G$  but since  $d_W(\mathcal{D}_{real}, \mathcal{D}_G) > 1$ .

---

<sup>14</sup>Note that this *does not* contradict Arjovsky since they actually use a parameterized surrogate rather than the actual Wasserstein distance (i.e. a neural net with finite parameters).

## Definition ( $\mathcal{F}$ -distance)

Let  $\mathcal{F}$  be a class of functions from  $\mathbb{R}^d$  to  $[0, 1]$  such that if  $f \in \mathcal{F} \implies 1 - f \in \mathcal{F}$ , and let  $\phi$  be a concave measuring function. Then the  $\mathcal{F}$ -divergence with respect to  $\phi$  between distributions  $\mu$  and  $\nu$  supported on  $\mathbb{R}^d$  is

$$d_{\mathcal{F}, \phi} = \sup_{D \in \mathcal{F}} \mathbb{E}_{x \sim \mu}[\phi(D(x))] + \mathbb{E}_{x \sim \nu}[\phi(1 - D(x))] - 2\phi(1/2)$$

This is the same as the definition we saw before, but makes clear that the choice of discriminator class ( $\mathcal{F}$ ) is a *crucial* part of the definition

- ▶ If  $\phi(t) = \log t$  and  $\mathcal{F}$  is all functions from  $\mathbb{R}^d \rightarrow [0, 1]$ , then  $d_{\mathcal{F}, \phi}$  is the same as JS divergence
- ▶ If  $\phi(t) = t$  and  $\mathcal{F}$  is the set of all 1-Lipschitz functions from  $\mathbb{R}^d \rightarrow [0, 1]$ ,  $d_{\mathcal{F}, \phi}$  is the Wasserstein distance.
- ▶ if  $\mathcal{F}$  is the set of neural networks and  $\phi(t) = \log t$ , we get the original stated GAN objective
- ▶ Other choices give other varieties of GAN (EBGANS - total variation, MMD and GMMNs - bounded  $\infty$ -norm over some Reproducing Kernel Hilbert Space, etc.)
- ▶ In practice, though,  $\mathcal{F}$  is really the class of neural networks *with some bound  $p$  on the number of parameters*, *not* the general class of neural networks or bounded norm functions

Setting: Suppose  $\phi$  is bounded on  $[-\Delta, \Delta]$  and  $L_\phi$  is Lipschitz. Let  $\mathcal{F} = \{D_\nu, \nu \in \mathcal{V}\}$  be a class of discriminators that is  $L$ -Lipschitz with respect to parameters  $\nu$ , and  $p$  is the number of parameters in  $\nu$ .

## Theorem (Weak Generalization)

Let  $\mu, \nu$  be two distributions and  $\hat{\mu}, \hat{\nu}$  be empirical versions with at least  $m$  samples each. Then there is a constant  $c$  such that if  $m \geq \frac{cp\Delta^2 \log(LL_\phi p/\epsilon)}{\epsilon^2}$ , then with probability  $1 - e^{-p}$ , we have that

$$|d_{\mathcal{F},\phi}(\hat{\mu}, \hat{\nu}) - d_{\mathcal{F},\phi}(\mu, \nu)| \leq \epsilon$$

In other words, the *neural network* distance *does* generalize...

... but this theorem is actually *bad news* for diversity, since we can take  $\nu = \hat{\mu}$  to get that

$$|d_{F,\phi}(\mu, \hat{\mu})| \leq \epsilon$$

with  $\tilde{O}(p/\epsilon^2)$  samples.

In other words, *a discriminator with bounded capacity can't distinguish between a distribution  $\mu$  and a discrete distribution with  $\tilde{O}(p/\epsilon^2)$  sample size!*



- ▶ Didn't Goodfellow prove that GANs' distributions converge to the true distribution?
- ▶ Well, yes.... But, this was an asymptotic result.
- ▶ What these results show is that while we can guarantee NN-distance-generalization after polynomially many samples, we can't differentiate a between the true distribution and otherwise in polynomially many samples.
- ▶ In other words, to get finite-sample generalization bounds (in the worst case) from the GAN framework with bounded capacity discriminators, we could need *exponentially* many samples

## A deeper dive into generalizability and discriminatory capacity

Recall that there are various notions of convergence for random variables - one that is useful is called convergence *weak* convergence, which is defined by:

$$\mu_n \xrightarrow{w} \mu \iff \mathbb{E}_{\mu_n} f \rightarrow \mathbb{E}_{\mu} f$$

for all bounded continuous functions  $f$ . This is equivalent, by the Portmanteau theorem, to a more natural definition of convergence in distribution, which is

$$\mu_n \xrightarrow{D} \mu \iff \mu_n(A) \rightarrow \mu(A)$$

for every measurable event  $A$ .

- ▶ Recall that we previously defined several measures of distance between probability distribution. Fixing a distance  $d$  over the space of probability measures creates a metric space.
- ▶ Recall that we can view GANs as minimizing notions of distance between the generated and real probability distribution.
- ▶ We might hope that convergence of a distance measure implies convergence *in distribution* - if so, we say that a distance measure *metrizes* convergence in distribution.
- ▶ We might even hope that all our favorite distances do that.

It turns out that the distances we talked about have different strengths. For example, consider  $Z \sim U[0, 1]$ , and  $\mathbb{P}_0$  the distribution of  $(0, Z) \in \mathbb{R}^2$  (i.e. 0 on the x axis and random  $Z$  on the y-axis). Let  $g_\theta(z) = (\theta, z)$ . Then:

$$W(\mathbb{P}_0, \mathbb{P}_\theta) = |\theta|$$

$$\delta(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} 1 & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}$$

$$JS(\mathbb{P}_0, \mathbb{P}_\theta) = \begin{cases} \log 2 & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}$$

$$KL(\mathbb{P}_\theta \parallel \mathbb{P}_0) = KL(\mathbb{P}_0 \parallel \mathbb{P}_\theta) = \begin{cases} \infty & \theta \neq 0 \\ 0 & \theta = 0 \end{cases}$$

That is, as  $\theta \rightarrow 0$  (and so  $\mathbb{P}_\theta \xrightarrow{D} \mathbb{P}_0$ ), only the  $W(\mathbb{P}_0, \mathbb{P}_\theta)$  converges.

## Definition

Let  $(X, || \cdot ||)$  be a metric space and  $\mathcal{F}$  be a set of functions on  $X$ . We say that the pair  $d_{\mathcal{F}}(\mu, \nu)$  and  $\mathcal{F}$  is *discriminative* if

$$d_{\mathcal{F}}(\mu, \nu) = 0 \iff \mu = \nu$$

for any Borel probability measures  $\mu, \nu$  on  $X$ .

Notice that this definition says that a set  $\mathcal{F}$  is discriminative if the statement that  $\mathbb{E}_{\mu}[f] = \mathbb{E}_{\nu}[f]$  for every  $f \in \mathcal{F}$  implies that  $\mu = \nu$ .

Suppose we have a function set  $\mathcal{F} \subset C_b(X)$  where  $C_b$  is the set of continuous bounded functions on  $X$ , and  $X$  is a metric space.

## Theorem

*Define*

$$\text{span}\mathcal{F} := \left\{ \alpha_0 + \sum_{i=1}^n \alpha_i f_i : \alpha_i \in \mathbb{R}, f_i \in \mathcal{F}, n \in \mathbb{N} \right\}$$

*If for any  $f \in C_b$  and  $\epsilon > 0$  there exists  $f_\epsilon \in \text{span}\mathcal{F}$  such that  $\|f - f_\epsilon\|_\infty \leq \epsilon$  (i.e.  $\text{span}\mathcal{F}$  is dense in  $C_b(X)$  under  $\|\cdot\|_\infty$ ), then  $d_{\mathcal{F}}(\mu, \nu)$  is discriminative as defined above.*

An equivalent way to say it is that  $C_b(X) \subseteq \text{cl}(\text{span}\mathcal{F})$  (and it turns out, if  $X$  is compact, then this condition is necessary if  $d_{\mathcal{F}}(\mu, \nu)$  is to be discriminative).

The above states that if  $C_b(X)$  is contained in  $cl(span\mathcal{F})$ , then  $d_{\mathcal{F}}$  is discriminative. But note that we can write any 1-layer neural network as the span of signal neuron neural networks.

That is, if we define

$$\mathcal{F}_{nn} := \{\sigma(w^{\top}x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$$

we have the following theorem:

**Theorem (Thm 1 in Leshno et. al. 1993)**

*Let  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$  be a continuous activation function,  $X \subset \mathbb{R}^d$  be a compact set, and  $\mathcal{F}_{nn}$  as above. Then  $span\mathcal{F}_{nn}$  is dense in  $C(X)$   $\iff \sigma$  is not a polynomial.*



We get thus get that the neural network distance is discriminative (even when the parameter restriction is so strong as to require just 1 neuron!). Again, in practice though, we won't have  $d_{\mathcal{F}}(\mu, \nu) = 0$  for finite samples. So we need the following theorem to make this result useful:

## Theorem






Let  $(X, d_X)$  be any metric space. If  $\text{span}\mathcal{F}$  is dense in  $C_b(X)$ , then  $\lim_{n \rightarrow \infty} d_{\mathcal{F}}(\mu, \nu_n) = 0 \implies \nu_n \xrightarrow{w} \mu$ . Moreover, if  $\exists C : \|f\|_{BL} \leq C \forall f \in \mathcal{F}$ , then<sup>15</sup>  $\nu_n \xrightarrow{w} \mu \implies \lim_{n \rightarrow \infty} d_{\mathcal{F}}(\mu, \nu_n) = 0$ .





In other words, we have the theorem we like, and a partial converse under the assumption that the parameter space for the neural nets is bounded, which is true in practice.

---

<sup>15</sup>where  $\|\cdot\|_{BL} = \max\{\|\cdot\|_{\infty}, \|\cdot\|_{Lip}\}$ . One layer deeper:  
 $\|\cdot\|_{Lip} := \sup\{|f(x) - f(y)|/||x - y|| : x, y \in X, x \neq y\}$ .

- ▶ For whatever discriminator  $\mathcal{F}$ , we want  $d_{\mathcal{F}}(\mu, \nu_m) \rightarrow 0$  if and only if  $\nu_m \xrightarrow{w} \mu$  for any probability measures  $\mu$  and  $\nu_m$ . That means that
- ▶  $\mathcal{F}$  should be large enough to make  $d_{\mathcal{F}}(\mu, \nu)$  discriminative and that  $d_{\mathcal{F}}(\mu, \nu) \rightarrow 0$  implies  $\nu_m$  converges to  $\mu$ . This makes the learning objective valid in that the true distribution is being reached.
- ▶ But  $\mathcal{F}$  should be small enough that  $\nu_m \xrightarrow{w} \mu$  implies  $d_{\mathcal{F}}(\mu, \nu_m)$  approaches to 0. This makes sure that train and test loss are similar and thus the algorithm is *generalizable*.
- ▶ Turns out that for the neural-net distance, generalizability depends on the Rademacher complexity of  $\mathcal{F}$ , and the complexity of  $\mathcal{G}$  is immaterial.
- ▶ On the other hand, for stronger distances like KL divergence, increasing the complexity of  $\mathcal{G}$  does lead to overfitting without increasing the capacity of  $\mathcal{F}$  to discriminate.

-  Arjovsky, Chintala, Bottou  
Wasserstein GAN
-  Arora, Ge, Lianga, Ma, Zhang  
Generalization and Equilibrium in Generative Adversarial Nets.
-  Arora, Risteski, Do GANs learn the distribution? Some theory and empirics.
-  Athanassia G. Bacharoglou  
Approximation of Probability Distributions by Convex Mixtures of Gaussian Measures.
-  Chen, Duan, Houthoof, Schulman, Sutskever, Abbeel  
InfoGAN: Interpretable Representatoin Learning by Information Maximizing Generative Adversarial Nets.

-  Feizi, Suh, Xia, Tse  
Understanding GANs: the LQG Setting.
-  Goodfellow, Pouget-Abadie, Mirz, Xu, Warde-Farley, Ozair, Courville, Bengio  
Generative Adversarial Nets.
-  Ghosh et. al.  
Bayesian Nonparametrics.
-  Zhang, Liu, Zhou, Xu, He  
On the Discrimination-Generalization Tradeoff in GANs.