WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points

Albert No [1] **TaeHo Yoon** [2] **Sehyun Kwon** [2] **Ernest K. Ryu** [2]

Abstract

Generative adversarial networks (GAN) are a widely used class of deep generative models, but their minimax training dynamics are not understood very well. In this work, we show that GANs with a 2-layer infinite-width generator and a 2-layer finite-width discriminator trained with stochastic gradient ascent-descent have no spurious stationary points. We then show that when the width of the generator is finite but wide, there are no spurious stationary points within a ball whose radius becomes arbitrarily large (to cover the entire parameter space) as the width goes to infinity.

1. Introduction

Generative adversarial networks (GAN) (Goodfellow et al., 2014), which learn a generative model mimicking the data distribution, have found a broad range of applications in machine learning. While supervised learning setups solve minimization problems in training, GANs solve *minimax* optimization problems. However, the minimax training dynamics of GANs are poorly understood. Empirically, training is tricky to tune, as reported in (Mescheder et al., 2018, Section 1) and (Goodfellow, 2016, Section 5.1). Theoretically, prior analyses of minimax training have established few guarantees.

In the supervised learning setup, the limit in which the deep neural networks' width is infinite has been utilized to analyze the training dynamics. Another line of work establishes guarantees showing that no spurious local minima exist, and such results suggest (but do not formally guarantee) that training converges to the global minimum despite the non-convexity.

In this work, we study a Wasserstein GAN (WGAN) (Arjovsky et al., 2017; Gulrajani et al., 2017) with an infinitely wide generator trained with stochastic gradient ascent-descent. Specifically, we show that a WGAN with a 2-layer generator and a 2-layer discriminator both with random features and sigmoidal[1] activation functions and with the width of the generator (but not the discriminator) being large or infinite has no spurious stationary points[2] when trained with stochastic gradient ascent-descent. The theoretical analysis utilizes ideas from universal approximation theory and random feature learning.

1.1. Prior work

The classical universal approximation theorem establishes that a 2-layer neural network with a sigmoidal activation function can approximate any continuous function when the hidden layer is sufficiently wide (Cybenko, 1989). This universality result was extended to broader classes of activation functions (Hornik, 1991; Leshno et al., 1993), and quantitative bounds on the width of such approximations were established (Pisier, 1980-1981; Barron, 1993; Jones, 1992). Random feature learning (Rahimi & Recht, 2007; 2008a;b) combines these ingredients into the following implementable algorithm: generate the hidden layer weights randomly and optimize the weights of the output layers while keeping the hidden layer weights fixed.

In recent years, there has been intense interest in the analysis of infinitely wide neural networks, primarily in the realm of supervised learning. In the "lazy training regime", infinitely wide neural networks behave as Gaussian processes at initialization (Neal, 1996; Lee et al., 2018) and are essentially linear in the parameters, but not the inputs, during training. The limiting linear network can be characterized with the neural tangent kernel (NTK) (Jacot et al., 2018; Du et al., 2019; Li & Liang, 2018).

In a different "mean-field regime", the training dynamics of infinitely wide 2-layer neural networks are characterized with a Wasserstein gradient flow. This idea was concurrently developed by several groups (Chizat & Bach, 2018; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Rotskoff et al., 2019; Sirignano & Spiliopoulos, 2020a;b). Specif-

[1]Department of Electronic and Electrical Engineering, Hongik University, Seoul, Korea [2]Department of Mathematical Sciences, Seoul National University, Seoul, Korea. Correspondence to: Ernest K. Ryu <ernestryu@snu.ac.kr>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]We say an activation function is *sigmoidal* if it satisfies assumptions (AG) and (AD), which we later state. The standard sigmoid and $\tanh$ activations functions are sigmoidal.

[2]A stationary point is spurious if it is not a global minimum.

ically relevant to GANs, this mean-field machinery was applied to study the dynamics of finding mixed Nash equilibria of zero-sum games (Domingo-Enrich et al., 2020). Finally, Geiger et al. (2020) provides a unification of the NTK and mean-field limits.

Another line of analysis in supervised learning establishes that no spurious local minima, non-global local minima, exist. The first results of this type were established for the matrix and tensor decomposition setups (Ge et al., 2016; 2017; Wu et al., 2018; Sanjabi et al., 2019). Later, these analyses were extended to neural networks through the notion of no spurious "basins" (Nguyen et al., 2019; Liang et al., 2018b;a; Li et al., 2018a; Sun, 2020; Sun et al., 2020b) and "mode connectivity" (Garipov et al., 2018; Kuditipudi et al., 2019; Shevchenko & Mondelli, 2020).

Prior works have established convergence guarantees for GANs. The work of (Lei et al., 2020; Hsieh et al., 2019; Domingo-Enrich et al., 2020; Feizi et al., 2020; Sun et al., 2020a) establish global convergence as described in Section 1.2. Cho & Suh (2019) establish that the solution to the Wasserstein GAN is equivalent to PCA in the setup of learning a Gaussian distribution but do not make explicit guarantees on the training dynamics. Sanjabi et al. (2018) use a maximization oracle on a regularized Wassertstein distance to obtain an algorithm converging to stationary points, but did not provide any results relating to global optimality.

Although we do not make the connection formal, there is a large body of work establishing convergence for non-convex optimization problems with no spurious local minima solved with gradient descent (Lee et al., 2016; 2019) and stochastic gradient descent (Ge et al., 2015; Jin et al., 2017). The implication of having no spurious stationary points is that stochastic gradient descent finds a global minimum.

1.2. Contribution

The key technical challenge of this work is the non-convexity of the loss function in the generator parameters, caused by the fact that the discriminator is nonlinear and non-convex in the input. Prior work avoided this difficulty by using a linear discriminator (Lei et al., 2020) or by lifting the generator into the space of probability measures (Hsieh et al., 2019; Sun et al., 2020a; Domingo-Enrich et al., 2020), also described as finding mixed Nash equilibria, but these are modifications not commonly used in the empirical training of GANs. Feizi et al. (2020) seems to be the only exception, as they establish convergence guarantees for a WGAN with a linear generator and quadratic discriminator, but their setup is restricted to learning Gaussian distributions. In contrast, we use a nonlinear discriminator and directly optimize the parameters without lifting to find mixed Nash equilibria (we find pure Nash equilibria), while using standard stochastic gradient ascent-descent.

To the best of our knowledge, our work is the first to use infinite-width analysis to establish theoretical guarantees for GANs with a nonlinear discriminator trained with stochastic gradient-type methods. Our proof technique, distinct from the NTK or mean-field techniques, utilizes universal approximation theory and random feature learning to establish that there are no spurious stationary points. The only other prior work to use infinite-width analysis to study GANs was presented in (Domingo-Enrich et al., 2020), where the mean-field limit was used to establish guarantees on finding mixed Nash equilibria.

We point out that considering the NTK or mean-field limits of the generator and/or the discriminator networks does not resolve the non-convexity of the loss in the generator parameters. We adopt the random feature learning setup, where the hidden layer features are fixed, and optimize only the output layers for both the generator and the discriminator. Doing so allows us to focus on the key challenge of establishing guarantees on the optimization landscape despite the non-convexity.

2. Problem setup

We consider a WGAN whose generator and the discriminator are two-layer networks as illustrated in Figure 1.

Let $X \in \mathbb{R}^n$ be a random vector with a true (target) distribution P_X. Let $Z \in \mathbb{R}^k$ be a continuous random vector from the latent space satisfying the following assumption.

(AL) The latent vector $Z \in \mathbb{R}^k$ has a Lipschitz continuous probability density function $q_Z(z)$ satisfying $q_Z(z) > 0$ for all $z \in \mathbb{R}^k$.

The standard Gaussian is a possible choice satisfying (AL).

2.1. Generator Class

Let $\mathcal{G} = \{\phi(\cdot;\kappa) \,|\, \kappa \in \mathbb{R}^p\}$, where $\phi(\cdot;\kappa)\colon \mathbb{R}^k \to \mathbb{R}^n$, be a collection of generator feature functions satisfying the following assumption.

(AG) All generator feature functions $\phi \in \mathcal{G}$ are of form $\phi(z;\kappa) = \sigma_{\mathrm{g}}(\kappa_w z + \kappa_b)$, where $\kappa = (\kappa_w, \kappa_b) \in \mathbb{R}^{n\times k} \times \mathbb{R}^n$, and $\sigma_{\mathrm{g}}\colon \mathbb{R} \to \mathbb{R}$ is a bounded continuous activation function satisfying $\lim_{r\to-\infty} \sigma_{\mathrm{g}}(r) < \lim_{r\to\infty} \sigma_{\mathrm{g}}(r)$. (So $p = nk + n$.)

Definition 1 (Generator class, finite width)**.** *Consider the generator feature functions $\phi_1, \dots, \phi_{N_g} \in \mathcal{G}$, where $1 \le N_g < \infty$. For $\theta \in \mathbb{R}^{N_g}$, let*

$$g_\theta(z) = \sum_{i=1}^{N_g} \theta_i \phi_i(z).$$

Write

$$\mathrm{span}(\{\phi_i\}_{i=1}^{N_g}) = \left\{ g_\theta \,|\, \theta \in \mathbb{R}^{N_g} \right\}$$

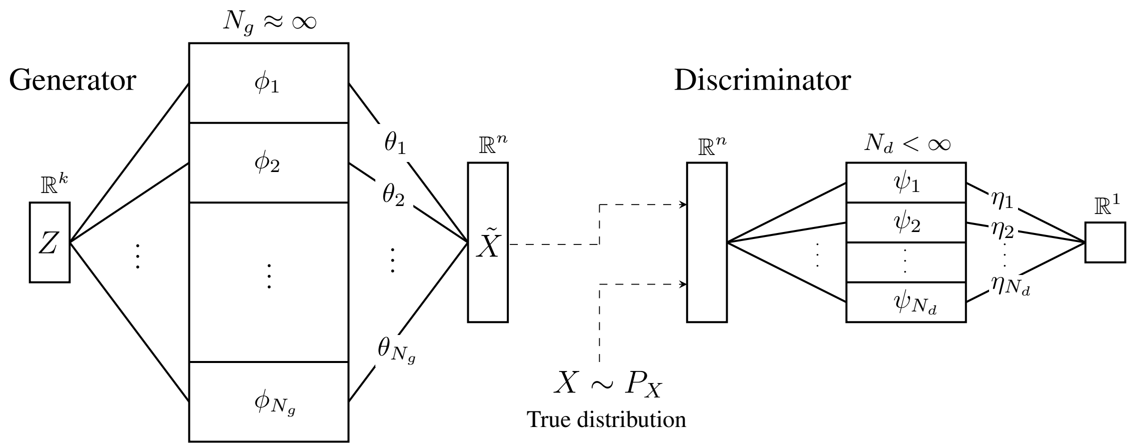


Figure 1. Illustration of the generator and discriminator architectures. Z represents the latent variable, $\tilde{X}$ the generated samples, and X the true samples. Respectively, $\{\phi_i\}_{i=1}^{N_g}$ and $\{\psi_j\}_{j=1}^{N_d}$ are the generator and discriminator (post-activation) feature functions. Respectively, $\{\theta_i\}_{i=1}^{N_g}$ and $\{\eta_j\}_{j=1}^{N_d}$ are the trainable parameters of the generator and discriminator networks.

for the class of generators constructed from the feature functions in $\{\phi_i\}_{i=1}^{N_g}$.

Note that there exists $\kappa_i \in \mathbb{R}^p$ such that $\phi_i(z) = \phi(z;\kappa_i)$ for $1 \le i \le N_g$. We can view the generator g_θ as a two-layer network, where $\phi(z;\kappa) = \sigma_{\rm g}(\kappa_w z + \kappa_b)$ represents the post-activation values of the hidden layer.

Definition 2 (Generator class, infinite width)**.** *For* $\theta \in \mathcal{M}(\mathbb{R}^p)$*, where* $\mathcal{M}(\mathbb{R}^p)$ *is the set of measures on* $\mathbb{R}^p$ *with finite total mass, let*

$$g_\theta(z) = \int \phi(z;\kappa)\, d\theta(\kappa).$$

Write

$$\overline{\operatorname{span}}(\mathcal{G}) = \{g_\theta(z) \,|\, \theta \in \mathcal{M}(\mathbb{R}^p)\}$$

for the class of infinite-width generators constructed from the feature functions in $\mathcal{G}$.

We assume the class of generator feature functions $\mathcal{G}$ satisfies the following universality property.

(Universal approximation property) For any function $f\colon \mathbb{R}^k \to \mathbb{R}^n$ such that $\mathbb{E}_Z\left[\|f(Z)\|_2\right] < \infty$ and $\varepsilon > 0$, there exists $\theta_\varepsilon \in \mathcal{M}(\mathbb{R}^p)$ such that

$$\mathbb{E}_Z\left[\|g_{\theta_\varepsilon}(Z) - f(Z)\|_2\right] < \varepsilon.$$

This assumption holds quite generally. In particular, the following lemma holds as a consequence of (Hornik, 1991).

Lemma 1. *(AG) implies* $\mathcal{G}$ *satisfies the (Universal approximation property).*

In functional analytical terms, (Universal approximation property) states that $\overline{\operatorname{span}}(\mathcal{G})$ is dense in $L^1(q_Z(z)\,dz;\mathbb{R}^n)$. Later in the proof of Theorem 4, we instead use the following dual characterization of denseness.

Lemma 2. *Assume (AL) and (Universal approximation property). If a bounded continuous function* $h\colon \mathbb{R}^k \to \mathbb{R}^n$ *satisfies*

$$\mathbb{E}_Z\left[\phi^\intercal(Z)h(Z)\right] = 0 \qquad \forall \phi \in \mathcal{G},$$

then $h(z) = 0$ *for all* $z \in \mathbb{R}^k$.

2.2. Discriminator Class

Let $\mathcal{D} = \{\psi_1, \ldots, \psi_{N_d}\}$ be a class of discriminator feature functions $\psi_j\colon \mathbb{R}^n \to \mathbb{R}$ for $1 \le j \le N_d$ satisfying the following assumption.

(AD) For all $1 \le j \le N_d$, the discriminator feature functions are of form $\psi_j(x) = \sigma(a_j^\intercal x + b_j)$ for some $a_j \in \mathbb{R}^n$ and $b_j \in \mathbb{R}$. The twice differentiable activation function σ satisfies $\sigma'(x) > 0$ for all $x \in \mathbb{R}$ and $\sup_{x\in\mathbb{R}} |\sigma(x)| + |\sigma'(x)| + |\sigma''(x)| < \infty$. The weights $a_1, \ldots, a_{N_d}$ and biases $b_1, \ldots, b_{N_d}$ are sampled (IID) from a distribution with a probability density function.

The sigmoid or $\tanh$ activation functions for σ and the standard Gaussian for the distribution of $a_1, \ldots, a_{N_d}$ and $b_1, \ldots, b_{N_d}$ are possible choices satisfying (AD). To clarify with measure-theoretic terms, we are assuming that $a_1, \ldots, a_{N_d}$ and $b_1, \ldots, b_{N_d}$ are sampled from a probability distribution that is absolutely continuous with respect to the Lebesgue measure.

Definition 3 (Discriminator Class)**.** *For* $\eta \in \mathbb{R}^{N_d}$*, let*

$$\Psi(x) = (\psi_1(x), \ldots, \psi_{N_d}(x)) \in \mathbb{R}^{N_d}$$

and

$$f_\eta(x) = \sum_{j=1}^{N_d} \eta_j \psi_j(x) = \eta^\intercal \Psi(x).$$

Write

$$\mathrm{span}(\mathcal{D}) = \left\{ f_\eta \mid \eta \in \mathbb{R}^{N_d} \right\}$$

for the class of discriminators constructed from the feature functions in $\mathcal{D}$.

In contrast with the generators, we only consider finite-width discriminators.

2.3. Adversarial training with stochastic gradients

Consider the loss

$$\begin{aligned} L(\theta, \eta) &= \mathbb{E}_X \left[f_\eta(X) \right] - \mathbb{E}_Z \left[f_\eta(g_\theta(Z)) \right] - \frac{1}{2}\|\eta\|_2^2 \\ &= \mathbb{E}_X \left[\eta^\intercal \Psi(X) \right] - \mathbb{E}_Z \left[\eta^\intercal \Psi(g_\theta(Z)) \right] - \frac{1}{2}\|\eta\|_2^2. \end{aligned}$$

This is a variant of the WGAN loss with the Lipschitz constraint on the discriminator replaced with an explicit regularizer. This loss and regularizer were also considered in (Lei et al., 2020).

We train the two networks adversarially by solving the minimax problem

$$\underset{\theta}{\text{minimize}} \ \underset{\eta}{\text{maximize}} \ L(\theta, \eta)$$

using stochastic gradient ascent-descent[3]

$$\begin{aligned} \gamma_\eta^t &= \Psi(X^t) - \Psi(g_\theta(Z_1^t)) - \eta^t \\ \eta^{t+1} &= \eta^t + \gamma_\eta^t \\ \gamma_\theta^t &= (D_\theta \Psi(g_\theta(Z_2^t)))^\intercal \eta^{t+1} \\ \theta^{t+1} &= \theta^t - \alpha \gamma_\theta^t \end{aligned}$$

for $t = 0, 1, \ldots$, where $X^t \sim P_X$, $Z_1^t \sim P_Z$, and $Z_2^t \sim P_Z$ are independent. We fix the maximization stepsize to 1 while letting the minimization stepsize be $\alpha > 0$. Note that γ_η^t and γ_θ^t are stochastic gradients in the sense that $\mathbb{E}\left[\gamma_\eta^t\right] = \nabla_\eta L(\theta^t, \eta^t)$ and $\mathbb{E}\left[\gamma_\theta^t\right] = \nabla_\theta L(\theta^t, \eta^{t+1})$. We can also form γ_η^t and γ_θ^t with batches. To clarify,

$$\begin{aligned} D_\theta \Psi(g_\theta(Z)) &= \begin{bmatrix} (\nabla_\theta (\psi_1(g_\theta(Z))))^\intercal \\ \vdots \\ (\nabla_\theta (\psi_{N_d}(g_\theta(Z))))^\intercal \end{bmatrix} \\ &= \begin{bmatrix} (\nabla_x \psi_1(g_\theta(Z)))^\intercal \\ \vdots \\ (\nabla_x \psi_{N_d}(g_\theta(Z)))^\intercal \end{bmatrix} \begin{bmatrix} \phi_1(Z) & \cdots & \phi_{N_g}(Z) \end{bmatrix}. \end{aligned}$$

[3] In the infinite-width case, where $\theta \in \mathcal{M}(\mathbb{R}^p)$, the stochastic gradient method we describe is not well defined as ∇_θ, if formally defined, is not an element of $\mathcal{M}(\mathbb{R}^p)$. For a rigorous treatment of analogs of gradient descent in $\mathcal{M}(\mathbb{R}^p)$, see (Mei et al., 2019; Chizat, 2021) and reference therein. In this work, we apply stochastic gradient ascent-descent only in the finite-width setup, but we analyze stationary points for both the finite and infinite setups.

The minimax problem is equivalent to the minimization problem

$$\inf_\theta \sup_\eta L(\theta, \eta) = \inf_\theta J(\theta),$$

where

$$\begin{aligned} J(\theta) &\triangleq \sup_\eta L(\theta, \eta) \\ &= \frac{1}{2} \left\| \mathbb{E}_X \left[\Psi(X) \right] - \mathbb{E}_Z \left[\Psi(g_\theta(Z)) \right] \right\|_2^2. \end{aligned}$$

Interestingly, stochastic gradient ascent-descent applied to $L(\theta, \eta)$ is equivalent to stochastic gradient descent applied to $J(\theta)$: eliminate the η-variable in the iteration to get

$$\theta^{t+1} = \theta^t - \alpha (D_\theta \Psi(g_\theta(Z_2^t)))^\intercal (\Psi(X^t) - \Psi(g_\theta(Z_1^t)))$$

and note

$$\begin{aligned} &\mathbb{E}\left[D_\theta \Psi(g_\theta(Z_2^t))^\intercal (\Psi(X^t) - \Psi(g_\theta(Z_1^t))) \right] \\ &\quad = \mathbb{E}_Z \left[(D_\theta \Psi(g_\theta(Z))^\intercal \right] (\mathbb{E}_X \left[\Psi(X) \right] - \mathbb{E}_Z \left[\Psi(g_\theta(Z)) \right]) \\ &\quad = \nabla_\theta J(\theta). \end{aligned}$$

In the following sections, we show that $J(\theta)$ has no spurious stationary points under suitable conditions.

Finally, we introduce the notation

$$r(\theta) = \mathbb{E}_X \left[\Psi(X) \right] - \mathbb{E}_Z \left[\Psi(g_\theta(Z)) \right],$$

i.e., $r_j(\theta) = \mathbb{E}_X \left[\psi_j(X) \right] - \mathbb{E}_Z \left[\psi_j(g_\theta(Z)) \right]$ for $1 \le j \le N_d$. This allows us to write $J(\theta) = \frac{1}{2}\|r(\theta)\|_2^2$.

3. Infinite-width generator

Consider a GAN with a two-layer *infinite-width* generator $g_\theta \in \overline{\mathrm{span}}(\mathcal{G})$ and a two-layer finite-width discriminator $f_\eta(x) \in \mathrm{span}(\mathcal{D})$. In this section, we show that under suitable conditions, $J(\theta)$ has no spurious stationary points, i.e., a stationary point of $J(\theta)$ is necessarily a global minimum.

We say θ_s is a stationary point of J if $J(\theta_\mathrm{s} + \lambda\mu)$, as a function of $\lambda \in \mathbb{R}$, is differentiable and has zero gradient at $\lambda = 0$ for any $\mu \in \mathcal{M}(\mathbb{R}^p)$.

3.1. Small discriminator ($N_d \le n$)

We first consider the case where the discriminator has width $N_d \le n$. Consider the following condition.

(Jacobian kernel point condition) The Jacobian

$$D\Psi(x)^\intercal = \begin{bmatrix} \nabla\psi_1(x) & \nabla\psi_2(x) & \cdots & \nabla\psi_{N_d}(x) \end{bmatrix}$$

satisfies $\ker(D\Psi(x)^\intercal) = \{0\}$ for all $x \in \mathbb{R}^n$.

We can interpret (Jacobian kernel point condition) to imply that there is no redundancy in the discriminator feature functions $\psi_1, \ldots, \psi_{N_d}$. This condition holds quite generally when sigmoidal activation functions are used, as characterized by the following lemma.

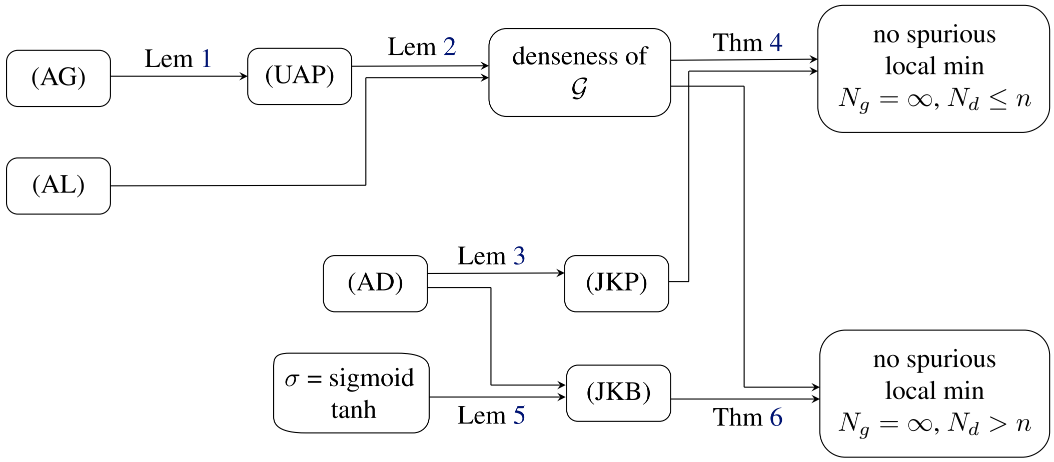


Figure 2. The diagrammatic summary of proofs in Section 3. We denote (Universal approximation property) by (UAP), (Jacobian kernel point condition) by (JKP), and (Jacobian kernel ball condition) by (JKB).

Lemma 3. *(AD) implies (Jacobian kernel point condition) with probability* 1.

Proof. Since $\nabla\psi_j(x)=a_j\sigma'(a_j^\intercal x+b_j)$, we have

$$D\Psi(x)^\intercal=\begin{bmatrix}\nabla\psi_1(x) & \cdots & \nabla\psi_{N_d}(x)\end{bmatrix}$$
$$=\begin{bmatrix}a_1 & \cdots & a_{N_d}\end{bmatrix}\operatorname{diag}(\sigma'(a_1^\intercal x+b_1),\dots,\sigma'(a_{N_d}^\intercal x+b_{N_d})).$$

By (AD), $\begin{bmatrix}a_1 & \cdots & a_{N_d}\end{bmatrix}$ has full rank with probability 1, and therefore $D\Psi(x)^\intercal$ has full rank. □

We are now ready to state and prove the first main result of this work: our GAN with an infinite-width generator has no spurious stationary points.

Theorem 4. *Assume (AL), (AG), and (AD). Then the following statement holds*[4] *with probability* 1*: any stationary point θ_s satisfies $J(\theta_s)=0$.*

Proof. Let θ_s be a stationary point of J. Then, for any $\mu\in\mathcal{M}(\mathbb{R}^p)$,

$$\frac{\partial}{\partial\lambda}J(\theta_s+\lambda\mu)\bigg|_{\lambda=0}=0.$$

Since $g_{\theta_s+\lambda\mu}=g_{\theta_s}+\lambda g_\mu$,

$$\begin{aligned}\frac{\partial}{\partial\lambda}J(\theta_s+\lambda\mu)&=\frac{\partial}{\partial\lambda}\frac{1}{2}\|r(\theta_s+\lambda\mu)\|_2^2\\&=-r(\theta_s+\lambda\mu)^\intercal\mathbb{E}_Z\left[\frac{\partial}{\partial\lambda}\Psi(g_{\theta_s+\lambda\mu}(Z))\right]\\&=-r(\theta_s+\lambda\mu)^\intercal\mathbb{E}_Z\left[D\Psi(g_{\theta_s+\lambda\mu}(Z))^\intercal g_\mu(Z)\right]\\&=-\mathbb{E}_Z\left[\sum_{j=1}^{N_d}r_j(\theta_s+\lambda\mu)\nabla\psi_j(g_{\theta_s+\lambda\mu}(Z))^\intercal g_\mu(Z)\right].\end{aligned}$$

Thus, for all $\mu\in\mathcal{M}(\mathbb{R}^p)$,

$$\mathbb{E}_Z\left[\sum_{j=1}^{N_d}r_j(\theta_s)\nabla\psi_j(g_{\theta_s}(Z))^\intercal g_\mu(Z)\right]=0.$$

By Lemmas 1 and 2,

$$\sum_{j=1}^{N_d}r_j(\theta_s)\nabla\psi_j(g_{\theta_s}(z))=0 \tag{1}$$

for all $z\in\mathbb{R}^k$. Thus,

$$\begin{bmatrix}\nabla\psi_1(g_{\theta_s}(z)) & \cdots & \nabla\psi_{N_d}(g_{\theta_s}(z))\end{bmatrix}\begin{bmatrix}r_1(\theta_s)\\ \vdots\\ r_{N_d}(\theta_s)\end{bmatrix}=0.$$

By Lemma 3, the (Jacobian kernel point condition) holds with probability 1. Therefore, $r(\theta_s)=0$ and we conclude $J(\theta_s)=0$. □

3.2. Large discriminator $(n<N_d<\infty)$

Next, consider the case where the discriminator has width $N_d>n$. In the small discriminator case, we used the (Jacobian kernel point condition), which states $\operatorname{rank}(D\Psi(x)^\intercal)=$

[4]The randomness comes from the random generation of ψ_j's described in (AD) and is unrelated to randomness of SGD. Once ψ_j's have been generated and the (Jacobian kernel point condition) holds by Lemma 3, the conclusion of Theorem 4 holds without further probabilistic quantifiers.

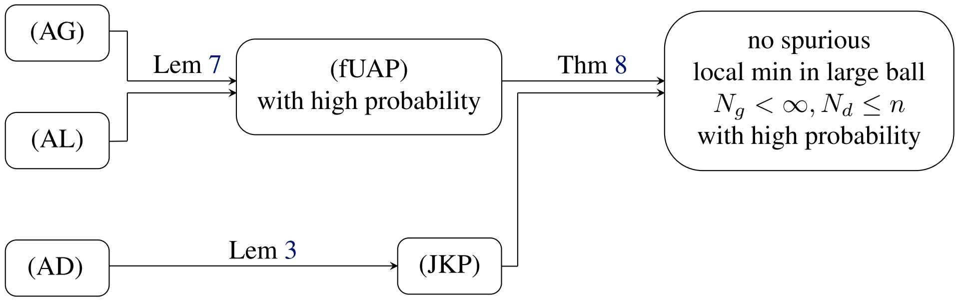


Figure 3. The diagrammatic summary of the proof in Section 4. We denote (Finite universal approximation property) by (fUAP), (Jacobian kernel point condition) by (JKP), and (Jacobian kernel ball condition) by (JKB).

N_d. However, this is not possible in the large discriminator case as $\operatorname{rank}(D\Psi(x)^\intercal) \leq n < N_d$. Therefore, we consider the following weaker condition.

(Jacobian kernel ball condition) For any open ball $B \subset \mathbb{R}^n$,

$$\bigcap_{x \in B} \ker(D\Psi(x)^\intercal) = \{0\}.$$

Since $\nabla_x(\eta^\intercal \Psi(x)) = D\Psi(x)^\intercal \eta$, the (Jacobian kernel ball condition) implies that $\eta^\intercal \Psi(x)$ with $\eta \neq 0$ is not a constant function within any open ball B, and we can interpret the condition to imply that there is no redundancy in the discriminator feature functions $\psi_1, \ldots, \psi_{N_d}$. The condition holds generically under mild conditions, as characterized by the following lemma.

Lemma 5. *Assume $\sigma \colon \mathbb{R} \to \mathbb{R}$ is the sigmoid or the $\tanh$ function. Then (AD) implies (Jacobian kernel ball condition) with probability 1.*

Proof outline of Lemma 5. The random generation of (AD) implies that with probability 1, all nonzero linear combinations of $\psi_1, \ldots, \psi_{N_d}$ are nonconstant, i.e., $f_\eta(x) = \eta^\intercal \Psi(x)$ with $\eta \neq 0$ is not globally constant (Sussmann, 1992, Lemma 1). Since σ is an analytic function, this implies f_η with $\eta \neq 0$ is not constant within any open ball B. So $\nabla_x f_\eta(x) = D\Psi(x)^\intercal \eta$ is not identically zero in B and we conclude $\eta \notin \bigcap_{x \in B} \ker(D\Psi(x)^\intercal)$ for any $\eta \neq 0$. □

We are now ready to state and prove the second main result of this work.

Theorem 6. *Assume (AL), (AG), and (AD). Then the following statement holds*[5] *with probability 1: for any stationary point θ_s, if the range of $g_{\theta_\mathrm{s}}(Z)$ contains an open-ball in $\mathbb{R}^n$, then $J(\theta_\mathrm{s}) = 0$.*

[5] Once ψ_j's have been generated and the (Jacobian kernel ball condition) holds by Lemma 5, the conclusion of Theorem 6 holds without further probabilistic quantifiers.

Proof. Following the same steps as in the proof of Theorem 4, we arrive at (1), which we rewrite as

$$D\Psi(g_\theta(z))^\intercal r(\theta) = 0.$$

Thus, for all $z \in \mathbb{R}^k$,

$$r(\theta) \in \ker(D\Psi(g_\theta(z))^\intercal).$$

Since the range of $X_\theta = g_\theta(Z)$ contains an open ball, the (Jacobian kernel ball condition), which holds with probability 1 by Lemma 5, implies $r(\theta) = 0$. □

Theorem 6 implies that a stationary point may be a spurious stationary point only when the generator's output is degenerate. One can argue that when P_X, the target distribution of X, has full-dimensional support, the generator should not converge to a distribution with degenerate support. Indeed, this is what we observe in our experiments of Section 5.

4. Finite-width generator

Consider a GAN with a two-layer *finite-width* generator $g_\theta \in \operatorname{span}(\{\phi_i\}_{i=1}^{N_g})$ and a two-layer finite-width discriminator $f_\eta(x) \in \operatorname{span}(\mathcal{D})$. In this section, we show that $J(\theta)$ has no spurious stationary points within a ball whose radius becomes arbitrarily large (to cover the entire parameter space) as the generator's width N_g goes to infinity.

The finite-width analysis relies on a finite version of the (Universal approximation property) that implies we can approximate a given function as a linear combination of $\{\phi_i\}_{i=1}^{N_g}$. Let $\delta^{(l)} : \mathbb{R}^k \to \mathbb{R}^n$ have the delta function on the l-th component and zero functions for all other components, i.e.,

$$\left[\delta^{(l)}(z)\right]_i = \begin{cases} 0 & \text{if } i \neq l \\ \delta(z) & \text{if } i = l \end{cases}$$

for $1 \leq l \leq n$.

(Finite universal approximation property) For a given $\varepsilon > 0$, there exists a large enough $N_g \in \mathbb{N}$ and

$\phi_1, \ldots, \phi_{N_g} \in \mathcal{G}$ such that there exists $\{\theta_i^{(\varepsilon,l)} \in \mathbb{R} \mid 1 \le i \le N_g, 1 \le l \le n\}$ satisfying

$$\left| \mathbb{E}_Z \left[\left(\sum_{i=1}^{N_g} \theta_i^{(\varepsilon,l)} \phi_i(Z) - \delta^{(l)}(Z) \right)^{\intercal} f(Z) \right] \right| < \varepsilon \sup_{z \in \mathbb{R}^k} \{ \|f(z)\|_2 + \|Df(z)\| \}$$

for all coordinates $l = 1, \ldots, n$, and for any continuously differentiable $f \colon \mathbb{R}^k \to \mathbb{R}^n$ such that $\sup_{z \in \mathbb{R}^k} \{\|f(z)\|_2 + \|Df(z)\|\} < \infty$.

This (Finite universal approximation property) holds with high probability when the width N_g is sufficiently large and the weights and biases of the generator feature functions $\phi_1, \ldots \phi_{N_g}$ are randomly generated.

Lemma 7. *Assume (AL) and (AG). Assume the first n parameters $\{\kappa_i\}_{i=1}^n$ are chosen so that $\{\phi_i\}_{i=1}^n$ are constant functions spanning the sample space $\mathbb{R}^n$. Assume the remaining parameters $\{\kappa_i\}_{i=n+1}^{N_g}$ are sampled (IID) from a probability distribution that has a continuous and strictly positive density function. Then for any $\varepsilon > 0$ and $\zeta > 0$, there exists large enough* [6] *N_g such that (Finite universal approximation property) with ε holds with probability at least $1 - \zeta$.*

Remember that the parameters define the generator feature functions through $\phi_i(x) = \phi_i(x; \kappa_i)$ for $1 \le i \le N_g$. By choosing the first n parameters in this way, we are effectively providing a trainable bias term in the output layer of the generator. Note that most universal approximation results consider the approximation of functions, while (Finite universal approximation property) requires the approximation of the delta function, which is not truly a function.

Proof outline of Lemma 7. Here, we illustrate the proof in the case of $n = 1$. The general $n \ge 1$ case requires similar reasoning but more complicated notation.

First, we define the smooth approximation of δ by

$$\tilde{\delta}^\varepsilon(z) = \frac{C}{\varepsilon^k} e^{-\|z/\varepsilon\|_2^2},$$

where C is a constant (depending on k but not ε) such that

$$\int_{\mathbb{R}^k} \tilde{\delta}^\varepsilon(z)\, dz = 1.$$

We argue that $\tilde{\delta}^\varepsilon(z) \approx \delta$ in the sense made precise in Lemma 11 of the appendix.

Next, we approximate $\tilde{\delta}^\varepsilon$ with the random feature functions. Using the arguments of (Barron, 1993, Theorem 2) and (Telgarsky, 2020, Section 4.2), we show that there exists a bounded density $m(\kappa)$ and $\kappa_1 \in \mathbb{R}^{k+1}$ such that $\phi_1 = \phi(z; \kappa_1)$ is a nonzero constant function and

$$\tilde{\delta}^\varepsilon(z) \approx \theta_1^\varepsilon \phi_1(z) + \int \phi(z; \kappa)\, m(\kappa)\, d\kappa$$

for some $\theta_1^\varepsilon \in \mathbb{R}$. For large $K > 0$,

$$\int \phi(z; \kappa)\, m(\kappa) d\kappa \approx \int \phi(z; \kappa)\, m(\kappa) \mathbf{1}_{\{\|\kappa\| \le K\}}(\kappa)\, d\kappa,$$

where $\mathbf{1}_{\{\|\kappa\| \le K\}}$ is the 0-1 indicator function. Write $p(\kappa)$ for the continuous and strictly positive density function of the distribution generating κ. Then $\sup_\kappa \{m(\kappa) \mathbf{1}_{\{\|\kappa\| \le K\}}(\kappa)/p(\kappa)\} < \infty$, and this allows us to use random feature learning arguments of (Rahimi & Recht, 2008b). By (Rahimi & Recht, 2008b, Lemma 1), there exists a large enough N_g such that there exist weights $\{\theta_i^\varepsilon\}_{i=2}^{N_g}$ such that

$$\sum_{i=2}^{N_g} \theta_i^\varepsilon \phi(z; \kappa_i) \approx \int \phi(z; \kappa)\, m(\kappa) \mathbf{1}_{\{\|\kappa\| \le K\}}(\kappa)\, d\kappa$$

with probability $1 - \zeta$. Finally, we complete the proof by chaining the $\approx$ steps. □

We are now ready to state and prove the third main result of this work: our GAN with an finite-width generator and small discriminator has no spurious stationary points within a large ball around the origin.

Theorem 8. *Let $N_d \le n$. Assume (AL), (AG), and (AD). Assume the generator feature functions are generated randomly as in Lemma 7. For any $C > 0$ and $\zeta > 0$, there exists a large enough $N_g \in \mathbb{N}$ such that the following statement holds with probability at least $1 - \zeta$: any stationary point $\theta_{\mathrm{s}} \in \mathbb{R}^{N_g}$ satisfying $\|\theta_{\mathrm{s}}\|_1 \le C$ is a global minimum.*[7]

Proof of Theorem 8. Since ϕ is bounded by (AG) and $\|\theta_{\mathrm{s}}\|_1$ is bounded, the output of $g_{\theta_{\mathrm{s}}}$ is also bounded, and

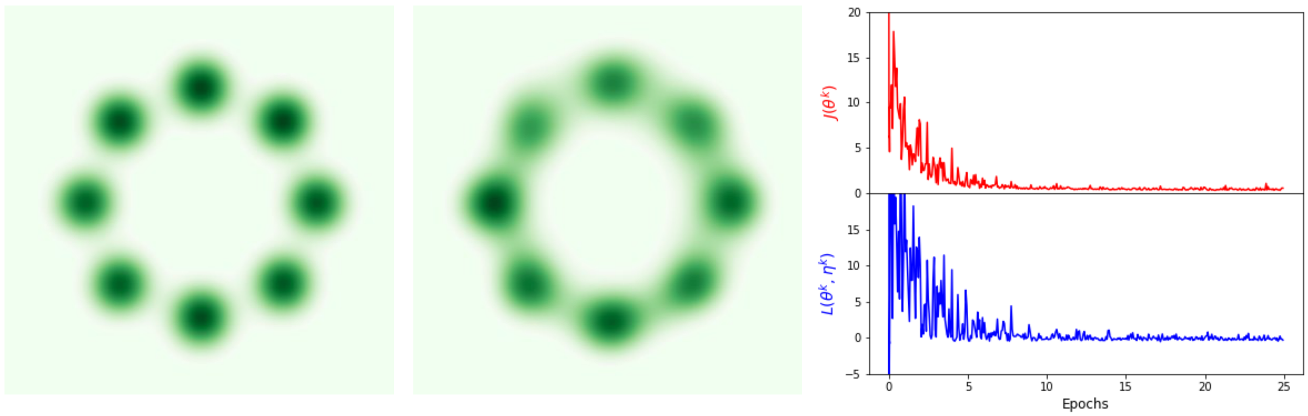
$$\sup_{\|\theta_{\mathrm{s}}\|_1 \le C} \|g_{\theta_{\mathrm{s}}}(0)\|_2 < \infty.$$

The (Jacobian kernel point condition), which holds with probability 1 by Lemma 3, implies

$$C_1 \triangleq \inf_{\|x\|_2 \le \left(\sup_{\|\theta_{\mathrm{s}}\|_1 \le C} \|g_{\theta_{\mathrm{s}}}(0)\|_2 \right)} \tau_{\min}(D\Psi(x)^\intercal) > 0,$$

[6] A quantitative bound on N_g can be established with careful bookkeeping. Specifically, using (10) of Appendix A.4, we can quantify N_g as a function of C_K. This C_K, which serves a similar role as the C of (Rahimi & Recht, 2007, Theorem 1), can also be quantified as a function of ϵ. However, the resulting bound is complicated and loose.

[7] The randomness comes from the random generation of ψ_j's and ϕ_i's described in (AD) and Lemma 7. Once the (Jacobian kernel point condition) and (Finite universal approximation property) holds (an event with probability at least $1 - \zeta$) the conclusion of Theorem 8 holds without further probabilistic quantifiers. Again, the size of N_g can be quantified through careful bookkeeping.



(a) Samples from true distribution P_X (b) Samples from generator $g_\theta(Z)$ (c) Convergence of the loss functions J and L

Figure 4. Samples and loss functions with a mixture of 8 Gaussians, $X \in \mathbb{R}^2$, $Z \in \mathbb{R}^2$, $N_g = 5,000$, and $N_d = 1,000$. The generator accurately learns the sampling distribution, and the loss functions converge to 0. The code is available at `https://github.com/sehyunkwon/Infinite-WGAN`.

where $\tau_{\min}$ denotes the N_d-th singular value. We use the fact that $\tau_{\min}(D\Psi(x)^\intercal)$ is a continuous function of x and the infimum over a compact set of a continuous positive function is positive. (We use $\tau_{\min}$ to denote the minimum singular value, rather than the standard $\sigma_{\min}$ to avoid confusion with the σ denoting the activation function.) By (AD),

$$C_2 \triangleq \max_{j=1,\dots,N_d} \sup_{x\in\mathbb{R}^n} \{\|\nabla\psi_j(x)\| + \|\nabla^2\psi_j(x)\|\} \in (0,\infty)$$

By Lemma 7, there exists a large enough N_g such that (Finite universal approximation property) with

$$\varepsilon = \frac{C_1 q_Z(0)}{2C_2 N_d n}$$

holds with probability $1-\zeta$.

Let θ_s be a stationary point satisfying $\|\theta_\mathrm{s}\|_1 \le C$. However, assume for contradiction that $J(\theta_\mathrm{s}) \neq 0$, i.e., $r(\theta_\mathrm{s}) \neq 0$. Then

$$\frac{\partial}{\partial\theta_i} J(\theta_\mathrm{s}) = \mathbb{E}_Z\left[\sum_{j=1}^{N_d} r_j(\theta_\mathrm{s})\nabla\psi_j(g_{\theta_\mathrm{s}}(Z))^\intercal \phi_i(Z)\right] = 0$$

for all $1 \le i \le N_g$. Define the normalized residual vector $\hat{r} = (1/\|r\|_2)r$, and write

$$\mathbb{E}_Z\left[\sum_{j=1}^{N_d} \hat{r}_j(\theta)\nabla\psi_j(g_\theta(Z))^\intercal \phi_i(Z)\right] = 0 \qquad (2)$$

for all $1 \le i \le N_g$.

Now consider

$$\begin{aligned}
&\left|\sum_{j=1}^{N_d} \hat{r}_j(\theta)\frac{\partial}{\partial x_l}\psi_j(g_\theta(0))q_Z(0)\right| \\
&= \left|\mathbb{E}_Z\left[\underbrace{\sum_{j=1}^{N_d} \hat{r}_j(\theta)\nabla\psi_j(g_\theta(Z))^\intercal}_{\triangleq f(Z)^\intercal} \delta^{(l)}(Z)\right]\right| \\
&= \left|\mathbb{E}_Z\left[f(Z)^\intercal\left(\delta^{(l)}(Z) - \sum_{i=1}^{N_g}\theta_i^{(l,\varepsilon)}\phi_i(z)\right)\right]\right| \\
&< \varepsilon \sup_{z\in\mathbb{R}^k}\{\|f(z)\| + \|Df(z)\|\} \\
&\le \varepsilon \sum_{j=1}^{N_d} |\hat{r}_j(\theta)| \sup_{x\in\mathbb{R}^n}\{\|\nabla\psi_j(x)\| + \|\nabla^2\psi_j(x)\|\} \\
&\le \varepsilon C_2 N_d,
\end{aligned}$$

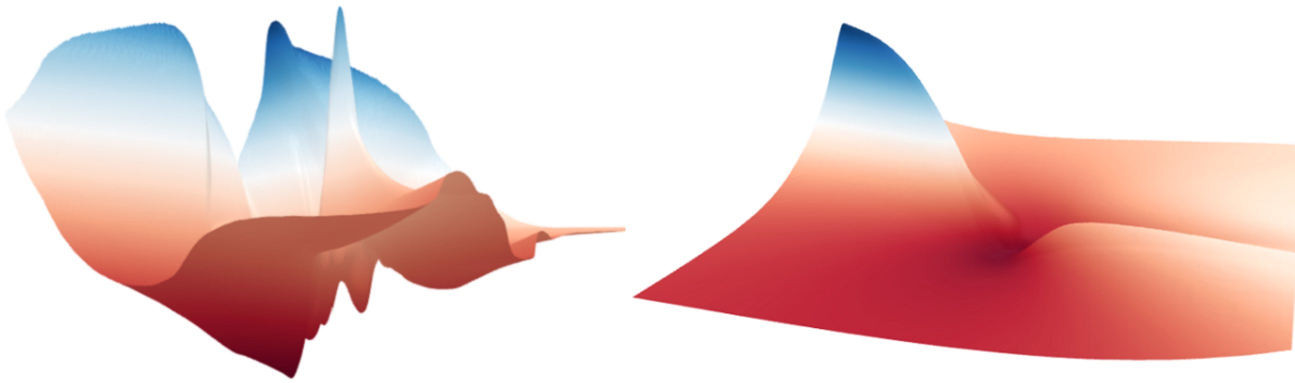
where the first equality follows from the definition of $\delta^{(l)}$, the second equality follows from (2), the first inequality follows from the the (Finite universal approximation property), the second inequality follows from the triangle inequality of the norm, and the third inequality follows from the definition of C_2 and the fact that the normalized residual satisfies $|\hat{r}_j(\theta)| \le 1$ for all j. By summing this result over $1 \le l \le n$ and using the bound $\|\cdot\|_2 \le \|\cdot\|_1$, we get

$$\|D\Psi(g_\theta(0))^\intercal \hat{r}(\theta)\|_2 < \varepsilon\frac{C_2 N_d n}{q_Z(0)}.$$

Finally, we arrive at

$$\begin{aligned}
C_1 = C_1\|\hat{r}(\theta)\|_2 &\le \|D\Psi(g_\theta(0))^\intercal \hat{r}(\theta)\|_2 \\
&< \varepsilon\frac{C_2 N_d n}{q_Z(0)} = \frac{C_1}{2},
\end{aligned}$$

which is a contradiction. □



(a) Loss landscape with $N_g = 2$ (b) Loss landscape with $N_g = 10$

Figure 5. The landscape of $J(\theta)$ for a mixture of two Gaussians with generator widths $N_g = 2$ and $N_g = 10$. The first $N_g = 2$ example has multiple non-global local minima. The second $N_g = 10$ example has no spurious stationary points despite the landscape being clearly non-convex. We provide corresponding contour plots in the appendix.

5. Experiments

Figure 6 presents an experiment with a mixture of 8 Gaussians and $N_g = 5,000$. The experiments demonstrate the sufficiency of two-layer networks with random features and that the training does not encounter local minima when N_g is large.

Figure 5 visualizes the loss landscape with generator widths $N_g = 2$ and $N_g = 10$. For the $N_g = 10$ case, the parameter space was projected down to a 2D space defined by random directions, as recommended by Li et al. (2018b). We observe the landscape becomes more favorable with larger width.

6. Conclusion

In this work, we presented an infinite-width analysis of a WGAN and established that no spurious stationary points exist under certain conditions.

At the same time, however, we point out that the infinite-width analysis does simplify away (hide) some finite phenomena. One such issue we encountered in our experiments was nearly vanishing gradients, which can occur despite the absence of spurious stationary points. A quantitative finite-width analysis establishing explicit bounds may provide an understanding and remedies to such issues and, therefore, is an interesting direction of future work.

Acknowledgements

AN was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2021R1F1A105956711). TY, SK, and EKR were supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2020R1F1A1A01072877], the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [No. 2017R1A5A1015626], by the New Faculty Startup Fund from Seoul National University, and by the AI Institute of Seoul National University (AIIS) through its AI Frontier Research Grant (No. 0670-20200015) in 2020. We thank Jisun Park for reviewing the manuscript and providing valuable feedback. Finally, we thank the anonymous referees for their suggestions on improving the clarity of the exposition and including the diagrammatic summaries of Figures 2 and 3.

References

Abramowitz, M. and Stegun, I. A. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. 9th edition, 1972.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. *ICML*, 2017.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Chizat, L. Sparse optimization on measures with overparameterized gradient descent. *Mathematical Programming*, 2021.

Chizat, L. and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *NeurIPS*, 2018.

Cho, J. and Suh, C. Wasserstein GAN can perform PCA. *Allerton Conference*, 2019.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Domingo-Enrich, Jelassi, Mensch, Rotskoff, and Bruna. A mean-field analysis of two-player zero-sum games. *NeurIPS*, 2020.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. Gradient descent provably optimizes over-parameterized neural networks. *ICLR*, 2019.

Feizi, S., Farnia, F., Ginart, T., and Tse, D. Understanding GANs in the LQG setting: Formulation, generalization and stability. *IEEE Journal on Selected Areas in Information Theory*, 1(1):304–311, 2020.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of DNNs. *NeurIPS*, 31, 2018.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points-online stochastic gradient for tensor decomposition. *COLT*, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *NeurIPS*, 2016.

Ge, R., Jin, C., and Zheng, Y. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *ICML*, 2017.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *NeurIPS*, 2014.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of Wasserstein GANs. *NeurIPS*, 2017.

Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.

Hsieh, Y.-P., Liu, C., and Cevher, V. Finding mixed Nash equilibria of generative adversarial networks. *ICML*, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *NeurIPS*, 2018.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. *ICML*, 2017.

Jones, L. K. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Annals of Statistics*, 20(1):608–613, 1992.

Kuditipudi, R., Wang, X., Lee, H., Zhang, Y., Li, Z., Hu, W., Ge, R., and Arora, S. Explaining landscape connectivity of low-cost solutions for multilayer nets. *NeurIPS*, 2019.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as Gaussian processes. *ICLR*, 2018.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. *COLT*, 2016.

Lee, J. D., Panageas, I., Piliouras, G., Simchowitz, M., Jordan, M. I., and Recht, B. First-order methods almost always avoid strict saddle points. *Mathematical Programming*, 176(1–2):311–337, 2019.

Lei, Q., Lee, J., Dimakis, A., and Daskalakis, C. SGD learns one-layer networks in WGANs. *ICML*, 2020.

Leshno, M., Lin, V. Y., Pinkus, A., and Schocken, S. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Li, D., Ding, T., and Sun, R. On the benefit of width for neural networks: Disappearance of basins. *arXiv:1812.11039*, 2018a.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *NeurIPS*, 2018b.

Li, Y. and Liang, Y. Learning overparameterized neural networks via stochastic gradient descent on structured data. *NeurIPS*, 2018.

Liang, S., Sun, R., Lee, J. D., and Srikant, R. Adding one neuron can eliminate all bad local minima. *NeurIPS*, 2018a.

Liang, S., Sun, R., Li, Y., and Srikant, R. Understanding the loss surface of neural networks for binary classification. *ICML*, 2018b.

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018.

Mei, S., Misiakiewicz, T., and Montanari, A. Mean-field theory of two-layers neural networks: Dimension-free bounds and kernel limit. *COLT*, (99), 2019.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? *ICML*, 2018.

Neal, R. M. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. 1996.

Nguyen, Q., Mukkamala, M. C., and Hein, M. On the loss landscape of a class of deep neural networks with no bad local valleys. *ICLR*, 2019.

Pisier, G. Remarques sur un résultat non publié de b. maurey. *Séminaire d'Analyse fonctionnelle (dit "Maurey-Schwartz")*, 1980-1981.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. *NeurIPS*, 2007.

Rahimi, A. and Recht, B. Uniform approximation of functions with random bases. *Allerton Conference*, 2008a.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. *NeurIPS*, 2008b.

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: Long time convergence and asymptotic error scaling of neural networks. *NeurIPS*, 2018.

Rotskoff, G., Jelassi, S., Bruna, J., and Vanden-Eijnden, E. Neuron birth-death dynamics accelerates gradient descent and converges asymptotically. *ICML*, 2019.

Sanjabi, M., Ba, J., Razaviyayn, M., and Lee, J. D. On the convergence and robustness of training GANs with regularized optimal transport. *NeurIPS*, 2018.

Sanjabi, M., Baharlouei, S., Razaviyayn, M., and Lee, J. D. When does non-orthogonal tensor decomposition have no spurious local minima? *arXiv:1911.09815*, 2019.

Shevchenko, A. and Mondelli, M. Landscape connectivity and dropout stability of SGD solutions for over-parameterized neural networks. *ICML*, 2020.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A law of large numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020a.

Sirignano, J. and Spiliopoulos, K. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and Their Applications*, 130(3):1820 – 1852, 2020b.

Sun, R. Optimization for deep learning: An overview. *Journal of the Operations Research Society of China*, 8(2): 249–294, 2020.

Sun, R., Fang, T., and Schwing, A. Towards a better global loss landscape of GANs. *NeurIPS*, 2020a.

Sun, R., Li, D., Liang, S., Ding, T., and Srikant, R. The global landscape of neural networks: An overview. *IEEE Signal Processing Magazine*, 37(5):95–108, 2020b.

Sussmann, H. J. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, 5(4):589–593, 1992.

Telgarsky, M. Deep learning theory lecture notes, Fall 2020.

Wu, C., Luo, J., and Lee, J. D. No spurious local minima in a two hidden unit ReLU network. *ICLR Workshop*, 2018.