

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). 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). Specifically relevant to GANs, this mean-field machinery was

---

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

1We 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.

2A stationary point is spurious if it is not a global minimum.
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; 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 Wasserstein 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 $G = \{ \phi(\cdot; \kappa) \mid \kappa \in \mathbb{R}^p \}$, where $\phi(\cdot; \kappa): \mathbb{R}^k \to \mathbb{R}^n$, be a collection of generator feature functions satisfying the following assumption.

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

Definition 1 (Generator class, finite width). Consider the generator feature functions $\phi_1, \ldots, \phi_N \in \mathcal{G}$, where $1 \leq N < \infty$. For $\theta \in \mathbb{R}^{Nz}$, let

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

Write

$$\text{span}(\{\phi_i\}_{i=1}^{Nz}) = \{g_\theta \mid \theta \in \mathbb{R}^{Nz}\}$$
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 \( 1 \leq i \leq N_g \). We can view the generator \( g_\theta \) as a two-layer network, where \( \phi(z; \kappa) = \sigma_g(\kappa w z + \kappa) \) 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

\[
\text{span}(\mathcal{G}) = \{ g_\theta(z) \mid \theta \in \mathcal{M}(\mathbb{R}^p) \}
\]

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 \( f : \mathbb{R}^k \rightarrow \mathbb{R}^n \) such that \( \mathbb{E}_Z \| f(Z) \|_2 < \infty \) and \( \varepsilon > 0 \), there exists \( \theta_\varepsilon \in \mathcal{M}(\mathbb{R}^p) \) such that

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

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 \( \text{span}(\mathcal{G}) \) is dense in \( L^1(\mathbb{P}_2(z) \, dz; \mathbb{R}^n) \). 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 \( h : \mathbb{R}^k \rightarrow \mathbb{R}^n \) satisfies

\[
\mathbb{E}_Z [\phi^T(\sigma(\sum_{j=1}^{N_d} \eta_j \psi_j(z)))] = 0 \quad \forall \phi \in \mathcal{G},
\]

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 \( \psi_j : \mathbb{R}^n \rightarrow \mathbb{R} \) for \( 1 \leq j \leq N_d \) satisfying the following assumption.

**(AD)** For all \( 1 \leq j \leq N_d \), the discriminator feature functions are of form \( \psi_j(x) = \sigma(a_j^T x + b_j) \) for some \( a_j \in \mathbb{R}^n \) and \( b_j \in \mathbb{R} \). The twice differentiable activation function \( \sigma \) 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 \( \sigma \) 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

\[
\eta \Psi(x) = \sum_{j=1}^{N_d} \eta_j \psi_j(x) = \eta^T \Psi(x).
\]
Write

\[
\text{span}(\mathcal{D}) = \{ f_\eta \mid \eta \in \mathbb{R}^{N_d} \}
\]

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

\[
L(\theta, \eta) = \mathbb{E}_X [f_\eta(X)] - \mathbb{E}_Z [\bar{g}_\theta(g_\theta(Z))] - \frac{1}{2} \|\eta\|_2^2
\]

\[
= \mathbb{E}_X [\eta^T \Psi(X)] - \mathbb{E}_Z [\eta^T \Psi(g_\theta(Z))] - \frac{1}{2} \|\eta\|_2^2.
\]

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

\[
\begin{align*}
\text{minimize } & L(\theta, \eta) \\
\text{maximize } & L(\theta, \eta)
\end{align*}
\]

using stochastic gradient ascent-descent\(^1\)

\[
\begin{align*}
\gamma^t_\eta &= \Psi(X^t) - \Psi(g_\theta(Z^t_1)) - \eta^t \\
\eta^{t+1} &= \eta^t + \gamma^t_\eta \\
\gamma^t_\theta &= (D_\theta \Psi(g_\theta(Z^t_1)))^T \eta^{t+1} \\
\theta^{t+1} &= \theta^t - \alpha \gamma^t_\theta
\end{align*}
\]

for \(t = 0, 1, \ldots\), where \(X^t \sim P_X\), \(Z^t_1 \sim P_Z\), and \(Z^t_2 \sim P_Z\) are independent. We fix the maximization stepsize to 1 while letting the minimization stepsize be \(\alpha > 0\). Note that \(\gamma^t_\eta\) and \(\gamma^t_\theta\) are stochastic gradients in the sense that 

\[
\mathbb{E} [\gamma^t_\eta] = \nabla_\eta L(\theta^t, \eta^t)\quad \text{and} \quad \mathbb{E} [\gamma^t_\theta] = \nabla_\theta L(\theta^t, \eta^{t+1}).
\]

We can also form \(\gamma^t_\eta\) and \(\gamma^t_\theta\) with batches. To clarify,

\[
D_\theta \Psi(g_\theta(Z)) = \left[
\begin{array}{c}
(\nabla_\theta (\psi_1(g_\theta(Z))))^T \\
\vdots \\
(\nabla_\theta (\psi_{N_d}(g_\theta(Z))))^T
\end{array}
\right] = \left[
\begin{array}{c}
(\nabla_x \psi_1(g_\theta(Z)))^T \\
\vdots \\
(\nabla_x \psi_{N_d}(g_\theta(Z)))^T
\end{array}
\right] \left[
\begin{array}{c}
\phi_1(Z) \\
\vdots \\
\phi_{N_d}(Z)
\end{array}
\right].
\]

The minimax problem is equivalent to the minimization problem

\[
\inf_{\theta} \sup_{\eta} L(\theta, \eta) = \inf_{\theta} J(\theta),
\]

where

\[
J(\theta) = \sup_{\eta} L(\theta, \eta)
\]

\[
= \frac{1}{2} \|\mathbb{E}_X [\Psi(X)] - \mathbb{E}_Z [\Psi(g_\theta(Z))]|^2.
\]

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

\[
\theta^{t+1} = \theta^t - \alpha (D_\theta \Psi(g_\theta(Z^t_1)))^T (\Psi(X^t) - \Psi(g_\theta(Z^t_1)))
\]

and note

\[
\mathbb{E} [D_\theta \Psi(g_\theta(Z^t_2))]^T (\Psi(X^t) - \Psi(g_\theta(Z^t_1)))
\]

\[
= \mathbb{E}_X [(D_\theta \Psi(g_\theta(Z))]^T (\mathbb{E}_X [\Psi(X)] - \mathbb{E}_Z [\Psi(g_\theta(Z)]))
\]

\[
= \nabla_\theta J(\theta).
\]

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

Finally, we introduce the notation

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

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

### 3. Infinite-width generator

Consider a GAN with a two-layer infinite-width generator \(g_\theta \in \text{span}(\mathcal{G})\) and a two-layer finite-width discriminator \(f_\eta \in \text{span}(\mathcal{D})\). 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 \(\theta_0\) is a stationary point of \(J\) if \(J(\theta_0 + \lambda \mu)\), 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 \((N_d \leq n)\)

We first consider the case where the discriminator has width \(N_d \leq n\). Consider the following condition.

**Jacobian kernel point condition** The Jacobian

\[
D \Psi(x)^T = \begin{bmatrix}
\nabla \psi_1(x) & \nabla \psi_2(x) & \ldots & \nabla \psi_{N_d}(x)
\end{bmatrix}
\]

satisfies \(\ker(D \Psi(x)^T) = \{0\}\) 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.
Lemma 3. (AD) implies (Jacobian kernel point condition) with probability 1.

Proof. Since $\nabla \psi_j(x) = a_j \sigma'(a_j^T x + b_j)$, we have

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

$$= [a_1 \cdots a_{N_d}] \text{diag}(\sigma'(a_1^T x + b_1), \ldots, \sigma'(a_{N_d}^T x + b_{N_d})).$$

By (AD), $[a_1 \cdots a_{N_d}]$ has full rank with probability 1, and therefore $D\Psi(x)^T$ has full rank.

We are now ready to state and prove the 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 $\theta_s$ satisfies $J(\theta_s) = 0$.

Proof. Let $\theta_s$ 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 $g_{\theta_s+\lambda \mu} = g_{\theta_s} + \lambda \mu$,

$$\frac{\partial}{\partial \lambda} J(\theta_s + \lambda \mu) = \frac{\partial}{\partial \lambda} 2 \|r(\theta_s + \lambda \mu)\|^2 = -r(\theta_s + \lambda \mu)^T \mathbb{E}_Z \left[ \frac{\partial}{\partial \lambda} \Psi(g_{\theta_s+\lambda \mu}(Z)) \right]$$

$$= -r(\theta_s + \lambda \mu)^T \mathbb{E}_Z [D\Psi(g_{\theta_s+\lambda \mu}(Z))^T g_{\mu}(Z)]$$

$$= -\mathbb{E}_Z \sum_{j=1}^{N_d} r_j(\theta_s + \lambda \mu) \nabla \psi_j(g_{\theta_s+\lambda \mu}(Z))^T g_{\mu}(Z).$$

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

$$\mathbb{E}_Z \sum_{j=1}^{N_d} r_j(\theta_s + \lambda \mu) \nabla \psi_j(g_{\theta_s}(Z))^T g_{\mu}(Z) = 0.$$

By Lemmas 1 and 2,

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

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

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

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

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

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

Since $\nabla_x(\eta^T \Psi(x)) = D\Psi(x)^T \eta$, the (Jacobian kernel ball condition) implies that $\eta^T \Psi(x)$ 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 $\sigma: \mathbb{R} \to \mathbb{R}$ 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., $f_\eta(x) = \eta^T \Psi(x)$ with $\eta \neq 0$ is not globally constant (Sussmann, 1992, Lemma 1). Since $\sigma$ is an analytic function, this implies $f_\eta$ with $\eta \neq 0$ is not constant within any open ball $B$. So $\nabla_x f_\eta(x) = D\Psi(x)^T \eta$ is not identically zero in $B$ and we conclude $\eta \notin \bigcap_{x \in B} \ker(D\Psi(x)^T)$ for any $\eta \neq 0$.

We are now ready to state and prove the 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 $\theta_s$, if the range of $g_{\theta_s}(Z)$ contains an open-ball in $\mathbb{R}^n$, then $J(\theta_s) = 0$.

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

$$D\Psi(g_{\theta}(z))^T r(\theta) = 0.$$

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

$^5$Once $\psi_j$’s have been generated and the (Jacobian kernel ball condition) holds by Lemma 5, the conclusion of Theorem 6 holds without further probabilistic quantifiers.
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 \( N_g \)-finite-width generator \( g_\theta \in \text{span}\{\phi_i\}_{i=1}^{N_g} \) and a two-layer finite-width discriminator \( f_\xi(x) \in \text{span}(D) \). 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.,

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

for \( 1 \leq l \leq n \). Although the delta “function” is not truly a function but rather a measure, this distinction is not needed as we only use the delta function as a notational shorthand.

(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 \( \{\theta_i^{(\varepsilon,l)} \in \mathbb{R} \mid 1 \leq i \leq N_g, 1 \leq l \leq n\} \) satisfying

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

for any continuously differentiable \( f : \mathbb{R}^k \to \mathbb{R}^n \) such that \( \sup_{z \in \mathbb{R}^k} \left\| f(z) \right\|_2 + \|Df(z)\|_2 < \infty \).

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 (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 \( N_g \) such that (Finite universal approximation property) with \( \varepsilon \) holds with probability at least \( 1 - \zeta \).

Remember that the parameters define the generator feature functions through \( \phi_1(x) = \phi_i(x; \kappa_i) \) for \( 1 \leq i \leq N_g \). 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 \( n \geq 1 \) case requires similar reasoning but more complicated notation.

First, we define the smooth approximation of \( \delta \) by

\[
\delta^\varepsilon(z) = \frac{C}{\varepsilon^2} e^{-\|z/\varepsilon\|^2_2},
\]

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

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

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

Next, we approximate \( \delta^\varepsilon \) 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

\[
\delta^\varepsilon(z) \approx \theta_i^\varepsilon \phi_i(z) + \int \phi(z; \kappa) \, m(\kappa) \, d\kappa
\]

for some \( \theta_i^\varepsilon \in \mathbb{R} \). For large \( K > 0 \),

\[
\int \phi(z; \kappa) \, m(\kappa) \, d\kappa \approx \int \phi(z; \kappa) \, m(\kappa) \, 1_{\{\|\kappa\| \leq K\}}(\kappa) \, d\kappa,
\]

where \( 1_{\{\|\kappa\| \leq K\}} \) is the 0-1 indicator function. Write \( p(\kappa) \) for the continuous and strictly positive density function of the distribution generating \( \kappa \). Then \( \sup_{\kappa} \{m(\kappa)1_{\{\|\kappa\| \leq K\}}(\kappa) / p(\kappa)\} < \infty \), 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

\[
\sum_{i=2}^{N_g} \theta_i^\varepsilon \phi(z; \kappa_i) \approx \int \phi(z; \kappa) \, m(\kappa) \, 1_{\{\|\kappa\| \leq K\}}(\kappa) \, d\kappa
\]
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$^6$ at least $1 - \zeta$: any stationary point $\theta_\ast \in \mathbb{R}^{N_g}$ satisfying $\|\theta_\ast\|_1 \leq C$ is a global minimum.

**Proof of Theorem 8.** Since $\phi$ is bounded by (AG) and $\|\theta_\ast\|_1$ is bounded, the output of $g_\theta$ is also bounded, and

$$\sup_{\|\theta_\ast\|_1 \leq C} \|g_\theta(0)\|_2 < \infty.$$ 

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

$$C_1 \triangleq \inf \left\{ \tau_{\min}(D\Psi(x)^T) > 0 \right\},$$

where $\tau_{\min}$ denotes the $N_d$-th singular value. We use the fact that $\tau_{\min}(D\Psi(x)^T)$ 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 $\sigma$ denoting the activation function.) By (AD),

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

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

$$\varepsilon = \frac{C_1qz(0)}{2C_2N_dn}$$

holds with probability $1 - \zeta$.

Let $\theta_\ast$ be a stationary point satisfying $\|\theta_\ast\|_1 \leq C$. However, assume for contradiction that $J(\theta_\ast) \neq 0$, i.e., $r(\theta_\ast) \neq 0$. Then

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

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

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

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

---

**Figure 2.** Samples and loss functions with a mixture of 8 Gaussians, $X \in \mathbb{R}^2$, $Z \in \mathbb{R}^2$, $N_a = 5,000$, 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.
WGAN with an Infinitely Wide Generator Has No Spurious Stationary Points

Figure 3. 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.

Now consider

$$\sum_{j=1}^{N_d} \hat{r}_j(\theta) \tilde{\partial}_{x_j} \psi_j(g_\theta(0)) q_Z(0)$$

which is a contradiction.

Finally, we arrive at

$$\|D\Psi(g_\theta(0))^T \hat{r}(\theta)\|_2 < \varepsilon \frac{C_2 N_d n}{q_Z(0)} = \frac{C_1}{2}$$

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 $|\hat{r}_j(\theta)| \leq 1$ for all $j$. By summing this result over $1 \leq l \leq n$ and using the bound $\|\cdot\|_2 \leq \|\cdot\|_1$, we get

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

5. Experiments

Figure 4 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 3 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.
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 over-parameterized 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.
WGaN with an Infinitely Wide Generator Has No Spurious Stationary Points

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. mau- rey. 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.
A. Omitted proofs

A.1. Proof of Lemma 1

Theorem 9 ((Hornik, 1991, Theorem 1)). Let $\sigma : \mathbb{R} \to \mathbb{R}$ be bounded and nonconstant and $P \in \mathcal{M}(\mathbb{R}^k)$ be a finite measure. Then for any $f \in L^1(P)$ and $\varepsilon > 0$, there exists $N \in \mathbb{N}$ and $\{((\theta_i, a_i), b_i) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}\}_{i=1}^N$ such that

$$\int_{\mathbb{R}^k} \left| \sum_{i=1}^N \theta_i \sigma(a_i^T z - b_i) - f(z) \right| dP(z) < \varepsilon.$$ 

To clarify, $f : \mathbb{R}^k \to \mathbb{R}$ in (Hornik, 1991, Theorem 1).

Proof of Lemma 1. Let $f : \mathbb{R}^k \to \mathbb{R}^n$ such that $\mathbb{E}_z[\|f(Z)\|_2] < \infty$. By (AG), $\sigma_{\ell_2}$ is a bounded nonconstant function. For $l = 1, \ldots, n$, Theorem 9 provides us with $N_l \in \mathbb{N}$ and $\{((\theta_i, a_i), b_i) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}\}_{i=1}^{N_l}$ such that

$$h_l(z) = \sum_{i=1}^{N_l} \theta_i \sigma_{\ell_2}(a_i^T z - b_i)$$

satisfies

$$\int_{\mathbb{R}^k} |h_l(z) - f_l(z)| q_z(z) dz < \frac{\varepsilon}{2^n},$$

where $f_l(z)$ is the $l$-th coordinate of $f(z) \in \mathbb{R}^n$ for $l = 1, \ldots, n$. Let $\ell_g = \lim_{r \to -\infty} \sigma_{\ell_2}(r)$. Let

$$A_i^{(l)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ (a_i^{(l)})^T \\ 0 \end{bmatrix} \quad \text{← l-th row,} \quad b_i^{(l),r} = \begin{bmatrix} r \\ \vdots \\ r \end{bmatrix}, \quad e_{-l} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \quad \text{← on l-th coordinates}$$

and

$$\tilde{f}^{(l),r}(z) = - \sum_{i=1}^{N_l} \theta_i^{(l)} \sigma_{\ell_2}(r) e_{-l} + \sum_{i=1}^{N_l} \theta_i^{(l)} \sigma_{\ell_2}(A_i^{(l)} z - b_i^{(l),r}).$$

Then, for each $l = 1, \ldots, n$, we have $\tilde{f}^{(l),r} = \sum_{i=1}^{N_l} \theta_i^{(l)} (\sigma_{\ell_2}(r) - \ell_g) \to 0$ as $r \to -\infty$ if $j \neq l$, while $\tilde{f}^{(l),r} = h_l(z)$. Because $\sigma_{\ell_2}$ is bounded, by Lebesgue’s dominated convergence theorem, we obtain

$$\lim_{r \to -\infty} \int_{\mathbb{R}^k} \left\| \begin{bmatrix} h_1(z) \\ \vdots \\ h_n(z) \end{bmatrix} - \sum_{l=1}^n \tilde{f}^{(l),r}(z) \right\|_1 q_z(z) dz = 0.$$ 

Therefore, there exists a large enough $r_{\text{big}} > 0$ such that

$$\int_{\mathbb{R}^k} \left\| \begin{bmatrix} h_1(z) \\ \vdots \\ h_n(z) \end{bmatrix} - \sum_{l=1}^n \tilde{f}^{(l),r_{\text{big}}}(z) \right\|_1 q_z(z) dz < \frac{\varepsilon}{2},$$

and we conclude with (3) that

$$\int_{\mathbb{R}^k} \left\| f(z) - \sum_{l=1}^n \tilde{f}^{(l),r_{\text{big}}}(z) \right\|_1 q_z(z) dz < \varepsilon.$$
Note that
\[ \sum_{i=1}^{n} \tilde{f}^{(i),r_{h_{\kappa_{i}}}}(z) \in \text{span}(\mathcal{G}) . \]

Therefore, using the bound \( \| \cdot \|_{2} \leq \| \cdot \|_{1} \), we get
\[ \int_{\mathbb{R}^{k}} \left\| f(z) - \sum_{i=1}^{n} \tilde{f}^{(i),r_{h_{\kappa_{i}}}}(z) \right\|_{2} q_{Z}(z) dz < \varepsilon . \]

\[ \square \]

**A.2. Proof of Lemma 2**

**Proof of Lemma 2.** Because \( h \) is bounded and \( q_{Z}(z) dz \) is a probability measure, we have \( \mathbb{E}_{Z} [ \| h(Z) \|_{2} ] < \infty \). Therefore, for any \( \varepsilon > 0 \), there exists \( \theta_{\varepsilon} \) such that \( \mathbb{E}_{Z} [ \| g_{\theta_{\varepsilon}}(Z) - h(Z) \|_{2} ] < \varepsilon \). Observe that
\[ \mathbb{E}_{Z} [ g_{\theta_{\varepsilon}}(Z) h(Z) ] = \int_{\mathbb{R}^{k}} \int_{\mathbb{R}^{p}} h^{T}(z) \phi(z; \kappa) d\theta_{\varepsilon}(\kappa) q_{Z}(z) dz \]
\[ = \int_{\mathbb{R}^{p}} \int_{\mathbb{R}^{k}} h^{T}(z) \phi(z; \kappa) q_{Z}(z) dz d\theta_{\varepsilon}(\kappa) \]
\[ = \int \mathbb{E}_{Z} [ h^{T}(Z) \phi(Z; \kappa) ] d\theta_{\varepsilon}(\kappa) = 0 . \]

Here the change in the order of integration is valid because \( \phi(z; \kappa) = \sigma_{\kappa_{w}}(z + \kappa_{b}) \leq \| \sigma_{\kappa_{w}} \|_{\infty} \) and the total mass of \( \theta_{\varepsilon} \) is finite, so that
\[ \int_{\mathbb{R}^{k}} \int_{\mathbb{R}^{p}} \| h^{T}(z) \phi(z; \kappa) \|_{2} d\theta_{\varepsilon}(\kappa) q_{Z}(z) dz \leq n \| h \|_{\infty} \| \sigma_{\kappa_{w}} \|_{\infty} \theta_{\varepsilon}(\mathbb{R}^{p}) < \infty . \]

To clarify, the \( \| \cdot \|_{\infty} \) for \( \| \sigma_{\kappa_{w}} \|_{\infty} \) is the standard supremum norm for \( L^{\infty} \) spaces while \( \| h \|_{\infty} = \max_{1 \leq i \leq n} \| h_{i} \|_{\infty} \) where \( h_{i}(z) \) is the \( i \)-th coordinate of \( h(z) \in \mathbb{R}^{n} \). Finally, we have
\[ \mathbb{E}_{Z} \left[ \| h(Z) \|_{2}^{2} \right] = \mathbb{E}_{Z} \left[ h^{T}(Z) (h(Z) - g_{\theta_{\varepsilon}}(Z)) \right] \leq \| h \|_{\infty} \mathbb{E}_{Z} \left[ \| h(Z) - g_{\theta_{\varepsilon}}(Z) \|_{1} \right] \]
\[ \leq \| h \|_{\infty} \mathbb{E}_{Z} \left[ \sqrt{n} \| h(Z) - g_{\theta_{\varepsilon}}(Z) \|_{2} \right] < \varepsilon \sqrt{n} \| h \|_{\infty} . \]

To clarify, \( \| h(Z) - g_{\theta_{\varepsilon}}(Z) \|_{1} \) denotes the \( \ell^{1} \) norm on the vector in \( \mathbb{R}^{n} \) for each \( z \). Now by letting \( \varepsilon \to 0 \), we have
\[ 0 = \mathbb{E}_{Z} \left[ \| h(Z) \|_{2}^{2} \right] = \int \| h(z) \|_{2}^{2} q_{Z}(z) dz . \]

Since \( q_{Z} \) is continuous and positive everywhere, we conclude that \( h(z) = 0 \) for all \( z \in \mathbb{R}^{k} \). \[ \square \]

**A.3. Proof of Lemma 5**

**Theorem 10 ((Sussmann, 1992, Lemma 1)).** Let \( \sigma = \text{tanh} \). Assume
\[ C_{0} + \sum_{j=1}^{N} \eta_{j} \sigma(a_{j}^T x + b_{j}) = C \]
for all \( x \in \mathbb{R}^{n} \), where \( \eta_{j} \neq 0 \) and \( a_{j} \neq 0 \) for \( 1 \leq j \leq N \). If there exists no distinct indices \( i \) and \( j \) such that \( (a_{i}, b_{i}) = \pm (a_{j}, b_{j}) \), then \( N = 0 \) (the sum vanishes) and \( C_{0} = C \).

**Proof of Lemma 5.** First consider the case where \( \sigma = \text{tanh} \). With probability 1, the condition of Theorem 10 holds, and
\[ F(x) \overset{\Delta}{=} \sum_{j=1}^{N_{d}} \eta_{j} \psi_{j}(x) \]
with $\eta \neq 0$ is not constant. Since $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is analytic on $\mathbb{R}$, it has a power series expansion

$$\sigma(t) = \sum_{\nu=0}^{\infty} s_{\nu} t^{\nu}.$$ 

Suppose that $0 \neq \eta \in \bigcap_{x \in \mathcal{B}} \ker(D\Psi(x)^T)$. Then

$$\sum_{j=1}^{N_d} \eta_j \nabla \psi_j(x) \equiv 0$$

for $x \in \mathcal{B}$, and

$$F(x) \Delta \sum_{j=1}^{N_d} \eta_j \psi_j(x) = \sum_{j=1}^{N_d} \eta_j \sum_{\nu=0}^{\infty} s_{\nu} (a_j^T x + b_j)^{\nu}$$

is constant for $x \in \mathcal{B}$. Fix any $x_0 \in \mathcal{B}$ and $u \in \mathbb{R}^n$. Let $\alpha_j = a_j^T u$ and $\beta_j = a_j^T x_0 + b_j$. Then for $x_u(t) \Delta x_0 + tu$,

$$F(x_u(t)) = \sum_{j=1}^{N_d} \eta_j \sum_{\nu=0}^{\infty} s_{\nu} (t \alpha_j + \beta_j)^{\nu}$$

$$= \sum_{m=0}^{\infty} \sum_{j=1}^{N_d} \eta_j s_{\nu} \left( \frac{\nu}{m} \right) \alpha_j^m \beta_j^m t^m$$

$$= \sum_{m=0}^{\infty} F_m t^m$$

is constant within $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. (Order of summations can be freely interchanged because power series for $\sigma$ are absolutely convergent for any choice of $t$.) But then $F_m$ must be zero for all $m \geq 1$, since $0 = \frac{d^m}{dt^m} \sum_{j=1}^{N_d} \eta_j \psi_j(x_u(0)) = m! F_m$. Therefore, in fact, $F(x_u(t)) \equiv F_0$ for all $t \in \mathbb{R}$, and $F_0 = F(x_0)$ does not depend on $u$. That is, $F$ is a constant function on $\mathbb{R}^n$. This implies that $\eta = 0$, which contradicts the assumption $\eta \neq 0$.

We extend the conclusion to the sigmoid function by noting that

$$\frac{1}{1 + e^{-r}} = \frac{\tanh(r/2) + 1}{2},$$

i.e., the sigmoid function is obtained by scaling the input of $\tanh$, adding a constant, and scaling the output. \hfill \Box

### A.4. Proof of Lemma 7

Recall that we defined

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

so that $\int_{\mathbb{R}^k} \tilde{\delta}(z) \, dz = 1$ for all $\varepsilon > 0$.

**Lemma 11.** There exists a constant $C_\delta$ depending only on $k$ but not on $\varepsilon > 0$ such that

$$\left| \mathbb{E}_Z \left[ \left( \tilde{\delta}(Z) - \delta(Z) \right) f(Z) \right] \right| < C_\delta \varepsilon \sup_{z \in \mathbb{R}^k} (\|f(z)\| + \|\nabla f(z)\|)$$

for all differentiable $f : \mathbb{R}^k \rightarrow \mathbb{R}$ such that $\sup_{z \in \mathbb{R}^k} (\|f(z)\| + \|\nabla f(z)\|) < \infty$. Here $\| \cdot \|$ denotes the operator norm, which coincides with the vector $\ell^2$ norm on $\mathbb{R}^k$. 
Proof. Let \( M = \|f\|_\infty \), \( L_f = \sup_{z \in \mathbb{R}^k} \|\nabla f(z)\| \) and let \( L_Z \) be the Lipschitz constant of \( f_Z(z) \). Then for any \( z \in \mathbb{R}^k \),

\[
|f(z)q_Z(z) - f(0)q_Z(0)| \leq |f(z)||q_Z(z) - q_z(0)| + |f(z) - f(0)|q_Z(0) \leq ML_Z\|z\| + L_f\|z\||q_Z(0).
\]

Integrating both sides over \( z \in \mathbb{R}^k \) with respect to \( \tilde{\delta}^\varepsilon(z) \) gives

\[
\int_{\mathbb{R}^k} |f(z)q_Z(z) - f(0)q_Z(0)| \tilde{\delta}^\varepsilon(z) \, dz \leq \int_{\mathbb{R}^k} (ML_Z + L_fq_Z(0))\|z\|\tilde{\delta}^\varepsilon(z) \, dz
\]

\[
= \int_{\mathbb{R}^k} (ML_Z + L_fq_Z(0))\frac{\pi^{-k/2}}{e^{\varepsilon\|z\|^2/2}} \, dz.
\]

Using change of variables, we rewrite and bound the last integral as

\[
(ML_Z + L_fq_Z(0))\pi^{-k/2}\varepsilon \int_{\mathbb{R}^k} \|z\|e^{-\|z\|^2/2} \, dz \leq \max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k} \|z\|e^{-\|z\|^2/2} \, dz\right) \varepsilon (M + L_f)
\]

\[
\leq 2\max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k} \|z\|e^{-\|z\|^2/2} \, dz\right) \varepsilon \sup_{z \in \mathbb{R}^k} (|f(z)| + \|Df(z)\|),
\]

which shows that

\[
\mathbb{E}_Z \left[\left(\tilde{\delta}^\varepsilon(Z) - \delta(Z)\right) f(Z)\right] \leq \int_{\mathbb{R}^k} |f(z)q_Z(z) - f(0)q_Z(0)| \tilde{\delta}^\varepsilon(z) \, dz
\]

\[
\leq C_\delta \varepsilon \sup_{z \in \mathbb{R}^k} (|f(z)| + \|Df(z)\|)
\]

where

\[
C_\delta = 2\max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k} \|z\|e^{-\|z\|^2/2} \, dz\right).
\]

\[
\square
\]

Lemma 12. (Abramowitz & Stegun, 1972, p. 302) Denote by \( \mathcal{F}[\cdot] \) be the Fourier transform operator. Then

\[
\mathcal{F}[\tilde{\delta}^\varepsilon]\omega = e^{-\pi^2\varepsilon^2\|\omega\|^2}.
\]

In particular, \( \mathcal{F}[\tilde{\delta}^\varepsilon]\omega \) is bounded, and

\[
\int_{\mathbb{R}^k} \mathcal{F}[\tilde{\delta}^\varepsilon]\omega \, d\omega < \infty,
\]

\[
\int_{\mathbb{R}^k} \|\omega\| \mathcal{F}[\tilde{\delta}^\varepsilon]\omega \, d\omega < \infty.
\]

We first provide a proof when \( n = 1 \), which conveys all important ideas of the proof. Although the general case involves significantly more complicated notations, it does not essentially differ from the simpler case.

Proof for the case \( n = 1 \).

Let \( \varepsilon > 0 \) be given.

**Step 1.** Using Lemma 11, approximate \( \delta(z) \) with \( \tilde{\delta}^\varepsilon(z) \).

**Step 2.** Approximate \( \tilde{\delta}^\varepsilon(z) \) with an infinite combination of functions in \( \mathcal{G} \).

Because both \( \tilde{\delta}^\varepsilon \) and \( \mathcal{F}[\tilde{\delta}^\varepsilon] \) are real-valued and positive, using the inverse Fourier transform, we can write

\[
\tilde{\delta}^\varepsilon(z) - \tilde{\delta}^\varepsilon(0) = \Re \int (e^{2\pi i \omega^T z} - 1) \mathcal{F}[\tilde{\delta}^\varepsilon](\omega) \, d\omega = \int (\cos (2\pi \omega^T z) - 1) \mathcal{F}[\tilde{\delta}^\varepsilon](\omega) \, d\omega
\]

(4)

for any \( z \in \mathbb{R}^k \). Note that by Lemma 12, the integral (4) is always well-defined.
Fix a large $R > 0$ satisfying
\[ \int_{\|z\| > R} q_z(z) \, dz < \frac{\pi^{k/2} e^{k+1}}{2}. \]

Following (Telgarsky, 2020, Section 4.2), for $\|z\| \leq R$, the cosine term in (4) can be rewritten as
\[ \cos(2\pi \omega^T z) - 1 = \int_0^{R\|\omega\|} -2\pi \sin(2\pi b) \, db = \int_0^R -2\pi \mathbf{1}_{[\omega^T z - b \geq 0]}(z) \sin(2\pi b) \, db + \int_0^R 2\pi \mathbf{1}_{[\omega^T z - b \leq 0]}(z) \sin(2\pi b) \, db. \quad (5) \]

Let $u_g = \lim_{r \to \infty} \sigma_g(r)$ and $\ell_g = \lim_{r \to -\infty} \sigma_g(r)$. Then by (AG), we have
\[ \mathbf{1}_{\{\tau \geq 0\}}(r) = \lim_{\tau \downarrow 0} \frac{1}{u_g - \ell_g} \left( \sigma_g \left( \frac{\tau}{\tau} \right) - \ell_g \right). \]

Hence we can approximate the step function terms in (5) using $\sigma_g$:
\[ \int_0^{R\|\omega\|} \lim_{\tau \downarrow 0} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( \frac{\omega^T z - b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, db + \int_0^{R\|\omega\|} \lim_{\tau \downarrow 0} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( -\frac{\omega^T z + b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, db. \quad (6) \]

Plugging (6) into (4), we obtain
\[ \delta^\tau(z) - \delta^\tau(0) = \int_0^{R\|\omega\|} \lim_{\tau \downarrow 0} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( \frac{\omega^T z - b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, F[\delta^\tau](\omega) \, d\omega + \int_0^{R\|\omega\|} \lim_{\tau \downarrow 0} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( -\frac{\omega^T z + b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, F[\delta^\tau](\omega) \, d\omega \quad (7) \]

for $\|z\| \leq R$.

Observe that because $\sigma_g$ is bounded and by Lemma 12, for any $\tau > 0$ and $z \in \mathbb{R}^k$,
\[ \int_0^{R\|\omega\|} \left| \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( \frac{\omega^T z - b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, F[\delta^\tau](\omega) \right| \, d\omega \leq \frac{2\pi (\|\sigma_g\|_{\infty} + \ell_g) R\|\omega\| \, F[\delta^\tau](\omega)}{u_g - \ell_g} < \infty. \]

Therefore, by Lebesgue’s dominated convergence theorem, we can freely change the order of integration and limit in (7). Using this fact, and applying change of variables, we can rewrite $\delta^\tau(z)$ as
\[ \delta^\tau(z) = \delta^\tau(0) + \lim_{\tau \downarrow 0} \int_0^{R\|\omega\|} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( \frac{\omega^T z - b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, F[\delta^\tau](\omega) \, d\omega + \lim_{\tau \downarrow 0} \int_0^{R\|\omega\|} \frac{2\pi}{u_g - \ell_g} \left( \sigma_g \left( -\frac{\omega^T z + b}{\tau} \right) - \ell_g \right) \sin(2\pi b) \, F[\delta^\tau](\omega) \, d\omega \]
\[ = \theta_1^\tau \phi(z; \kappa_1) + \lim_{\tau \downarrow 0} \int_0^{2\pi \|\omega\|} \frac{2\pi}{u_g - \ell_g} \sigma_g (\omega^T z) \sin(-2\pi \tau b) \, F[\delta^\tau](\tau \omega) \, d\omega + \lim_{\tau \downarrow 0} \int_0^{2\pi \|\omega\|} \frac{2\pi}{u_g - \ell_g} \sigma_g (\omega^T z) \sin(2\pi \tau b) \, F[\delta^\tau](\tau \omega) \, d\omega \]
\[ = \theta_1^\tau \phi(z; \kappa_1) + \lim_{\tau \downarrow 0} \int_{\mathbb{R}^k \times \mathbb{R}} \phi(z; \kappa) \tilde{m}_\tau(\kappa) \, d\kappa \]
for \(|z| \leq R\). We specify the notations that were newly introduced. First, we denoted \(\kappa = (\omega, b)\), so that \(\phi(z; \kappa) = \sigma_g(\omega^T z + b)\) (note that because we have assumed \(n = 1\), the generator parameter has dimension \(k + 1\)), and \(d\kappa\) is the Lebesgue measure on \(\mathbb{R}^k \times \mathbb{R}\). Next, we set \(\kappa_1 = (0, b_1)\) with some fixed \(b_1 \in \mathbb{R}\) satisfying \(\phi(z; \kappa_1) \equiv \sigma_g(b_1) \neq 0\) and

\[
\theta_1^\epsilon = \frac{1}{\sigma_g(b_1)} \left( \delta^\epsilon(0) + \int_0^{|R||\omega|} \frac{2\pi \ell_g}{u_g - \ell_g} \sin(2\pi b) & \mathcal{F}[\delta^\epsilon](\omega) \, db \, d\omega - \int_0^{|R||\omega|} \frac{2\pi \ell_g}{u_g - \ell_g} \sin(2\pi b) \mathcal{F}[\delta^\epsilon](\omega) \, db \, d\omega \right) \in \mathbb{R}.
\]

Finally, we define the density function \(m_\tau(\kappa)\) as

\[
m_\tau(\kappa) = \frac{2\pi \tau^{k+1}}{u_g - \ell_g} \left( -\sin(-2\pi \tau b) \mathcal{F}[\delta^\tau](\tau \omega) \mathbf{1}_{\{R||\omega|| \leq b \leq 0\}}(\kappa) + \sin(2\pi \tau b) \mathcal{F}[\delta^\tau](-\tau \omega) \mathbf{1}_{\{-R||\omega|| \leq b \leq 0\}}(\kappa) \right)
\]

\[
= \frac{4\pi \tau^{k+1}}{u_g - \ell_g} e^{-\pi^2 \tau^2 ||\omega||^2} \sin(2\pi \tau b) \mathbf{1}_{\{R||\omega|| \leq b \leq 0\}}(\kappa),
\]

where we used Lemma 12 to obtain the second equality.

Now we bound the error in using the expression (5) in the case \(|z| > R\). Observe that

\[
\cos(2\pi \omega^T z) - 1 - \int_0^{|R||\omega||} -2\pi \mathbf{1}_{\{\omega^T z - b \geq 0\}}(z) \sin(2\pi b) \, db + \int_0^{|R||\omega||} 2\pi \mathbf{1}_{\{\omega^T z - b \leq 0\}}(z) \sin(2\pi b) \, db
\]

\[
= (\cos(2\pi \omega^T z) - \cos(2\pi R||\omega||)) \mathbf{1}_{\{\omega^T z > R||\omega||\}}(\omega),
\]

and thus

\[
\delta^\epsilon(z) - \theta_1^\epsilon \phi(z; \kappa_1) = \lim_{\tau \downarrow 0} \int_{\mathbb{R}^k \times \mathbb{R}} \phi(z; \kappa) m_\tau(\kappa) \, d\kappa = \int_{\{\omega^T z > R||\omega||\}} (\cos(2\pi \omega^T z) - \cos(2\pi R||\omega||)) \mathcal{F}[\delta^\epsilon](\omega) \, d\omega
\]

for all \(z \in \mathbb{R}^k\). The defining equation (8) shows that \(m_\tau\) is bounded and \(m_\tau \in L^1(d\kappa)\) with

\[
\int_{\mathbb{R}^k \times \mathbb{R}} |m_\tau(\kappa)| \, d\kappa \leq \frac{4\pi R}{u_g - \ell_g} \int_{\mathbb{R}^k} \tau^k ||\omega|| e^{-\pi^2 \tau^2 ||\omega||^2} \, d\omega = \frac{4\pi R}{u_g - \ell_g} \int_{\mathbb{R}^k} ||\omega|| e^{-\pi^2 \tau^2 ||\omega||^2} \, d\omega.
\]

Therefore, the family

\[
\left\{ \delta^\epsilon(z) - \theta_1^\epsilon \phi(z; \kappa_1) - \int_{\mathbb{R}^k \times \mathbb{R}} \phi(z; \kappa) m_\tau(\kappa) \, d\kappa \right\}_{\tau > 0}
\]

is uniformly bounded. Applying the dominated convergence theorem to the pointwise convergence result (9) with respect to the probability measure \(q_Z(z) \, dz\), we obtain

\[
\lim_{\tau \downarrow 0} \mathbb{E}_Z \left[ \delta^\epsilon(Z) - \theta_1^\epsilon \phi(Z; \kappa_1) - \int_{\mathbb{R}^k \times \mathbb{R}} \phi(Z; \kappa) m_\tau(\kappa) \, d\kappa \right] = \mathbb{E}_Z \left[ \int_{\{\omega^T Z > R||\omega||\}} (\cos(2\pi \omega^T Z) - \cos(2\pi R||\omega||)) \mathcal{F}[\delta^\epsilon](\omega) \, d\omega \right]
\]

\[
\leq \mathbb{E}_Z \left[ \int_{\{\omega^T Z > R||\omega||\}} 2 \mathcal{F}[\delta^\epsilon](\omega) \, d\omega \right]
\]

\[
\leq \mathbb{E}_Z \left[ 1_{\{||Z|| > R\}}(Z) \int_{\mathbb{R}^k} 2 \mathcal{F}[\delta^\epsilon](\omega) \, d\omega \right]
\]

\[
= \left( \int_{\{||z|| > R\}} q_Z(z) \, dz \right) \left( \int_{\mathbb{R}^k} 2 \mathcal{F}[\delta^\epsilon](\omega) \, d\omega \right) < \frac{\pi^{k/2} e^{k+1}}{2} \frac{2}{\pi^{k/2} e^k} = \epsilon.
\]

**Step 3.** Approximate the integral over \(\mathbb{R}^k \times \mathbb{R}\) by an integral over a ball of finite radius.
We fix some $\tau = \tau(\varepsilon)$ satisfying $E_Z \left[ \hat{\delta}^\tau(Z) - \theta_1^\tau \phi(Z; \kappa_1) - \int_{\mathbb{R}^k} \phi(Z; \kappa)m_\tau(\kappa) \, d\kappa \right] < 2\varepsilon$. Because $\sigma_g$ is bounded and $m_\tau \in L^1(d\kappa)$, there exists $K > 0$ large enough so that

$$\int_{\|\kappa\| > K} |m_\tau(\kappa)| \, d\kappa < \frac{\varepsilon}{\|\sigma_g\|_\infty}.$$  

Then for any bounded continuous function $f : \mathbb{R}^k \to \mathbb{R}$ we have

$$E_Z \left[ \hat{\delta}^\tau(Z)f(Z) - \theta_1^\tau \phi(Z; \kappa_1)f(Z) - \int_{\|\kappa\| \leq K} \phi(Z; \kappa)f(Z)m_\tau(\kappa) \, d\kappa \right]$$

$$\leq E_Z \left[ \hat{\delta}^\tau(Z)f(Z) - \theta_1^\tau \phi(Z; \kappa_1)f(Z) - \int_{\mathbb{R}^k} \phi(Z; \kappa)f(Z)m_\tau(\kappa) \, d\kappa \right]$$

$$+ E_Z \left[ \int_{\|\kappa\| > K} |\phi(Z; \kappa)f(Z)m_\tau(\kappa)| \, d\kappa \right]$$

$$\leq 2\varepsilon\|f\|_\infty + \|\sigma_g\|_\infty \|f\|_\infty E_Z \left[ \int_{\|\kappa\| > K} |m_\tau(\kappa)| \, d\kappa \right] \leq 3\varepsilon\|f\|_\infty.$$  

**Step 4.** Approximate the integral over a finite ball by a finite linear combination of random functions in $G$.

Define

$$m_{\tau,K}(\kappa) = \begin{cases} m_\tau(\kappa) & \text{if } \|\kappa\| \leq K, \\ 0 & \text{otherwise}. \end{cases}$$

Denote by $p(\kappa)$ the strictly positive continuous density function from which we randomly sample the generator parameters.

Note that we have

$$C_K \triangleq \sup_{\kappa} \left| \frac{m_{\tau,K}(\kappa)}{p(\kappa)} \right| < \infty$$

because $\|m_\tau\|_\infty < \infty$ and $1/p(\kappa)$ is bounded over a compact set.

Now, rewrite the integral from Step 3 as

$$\int_{\|\kappa\| \leq K} \phi(z; \kappa) \, m_\tau(\kappa) \, d\kappa = \int \phi(z; \kappa) \frac{m_{\tau,K}(\kappa)}{p(\kappa)} \, p(\kappa) \, d\kappa.$$  

We will show that if we sample $\kappa_2, \ldots, \kappa_{N_g}$ (IID) according to $p(\kappa)$, then for sufficiently large $N_g$, 

$$\int \phi(Z; \kappa) \frac{m_{\tau,K}(\kappa)}{p(\kappa)} \, p(\kappa) \, d\kappa \approx \frac{1}{N_g} \sum_{i=2}^{N_g} \phi(Z; \kappa_i) \frac{m_{\tau,K}(\kappa_i)}{p(\kappa_i)}$$

with high probability over $\kappa_2, \ldots, \kappa_{N_g}$. (The indexing begins with $i = 2$ because $\kappa_1$ is reserved for the constant function.) 

When we draw each $\kappa_i$, we are in fact sampling the corresponding function

$$h_i \triangleq \frac{m_{\tau,K}(\kappa_i)}{p(\kappa_i)} \phi(\cdot; \kappa_i) \in \mathcal{H} \triangleq L^2(q_Z(z) \, dz).$$

Indeed, $h_i$ are uniformly bounded with $\|h_i\|_\infty \leq \|\sigma_g\|_\infty C_K$ for all $i = 2, \ldots, N_g$, which implies $\|h_i\|_\mathcal{H} \leq \|\sigma_g\|_\infty C_K$. 

That is, $\frac{m_{\tau,K}(\kappa)}{p(\kappa)} \phi(\cdot; \kappa)$ is a bounded random variable with random realizations in $\mathcal{H}$. Also, we have

$$E_{\kappa \sim p(\kappa)} \left[ \frac{m_{\tau,K}(\kappa)}{p(\kappa)} \phi(\cdot; \kappa) \right] = \int \phi(\cdot; \kappa) \frac{m_{\tau,K}(\kappa)}{p(\kappa)} \, p(\kappa) \, d\kappa = \int_{\|\kappa\| \leq K} \phi(\cdot; \kappa) m_\tau(\kappa) \, d\kappa.$$
The crux of the general case is that \( \kappa \) cannot be sampled coordinate-wisely, but we must keep only one coordinate active, while suppressing the others. To achieve this, we simply accept \( \kappa \)'s whose rows are negligibly small except possibly for the \( l \)-th row. We express \( \kappa \in \mathbb{R}^{n \times k} \times \mathbb{R}^n \) in the form

\[
\kappa = \begin{bmatrix}
\kappa^{(1)} \\
\vdots \\
\kappa^{(n)}
\end{bmatrix} = \begin{bmatrix}
(\omega^{(1)})^\top & b^{(1)} \\
\vdots & \vdots \\
(\omega^{(n)})^\top & b^{(n)}
\end{bmatrix},
\]

where \( \omega^{(j)} \in \mathbb{R}^k, b^{(j)} \in \mathbb{R} \) and \( \kappa^{(j)} = (\omega^{(j)}, b^{(j)}) \) for \( j = 1, \ldots, n \).
We will show that for sufficiently small $\xi$ and some constant vector $\nu(\xi)$, consider the set $\{l / 2 \leq k \leq 1\}$ such that for some $\rho(l) \in \mathbb{R}$ and $K > 0$ large enough,

$$E_Z \left[ \left| \tilde{\delta}^l (Z) - \rho^{(l)} - \int_{\|\kappa^{(l)}\| \leq K} \sigma_g \left( (\omega_l)^T Z + b^{(l)} \right) m(\kappa^{(l)}) d\kappa^{(l)} \right| \right] < 3\varepsilon.$$

Note that we can bound

$$\left| \int_{\|\kappa^{(l)}\| \leq K} m(\kappa^{(l)}) d\kappa^{(l)} \right| = \left| \int_{\|\omega^{(l)}\| \leq K} \frac{2\varepsilon^{k+1}}{2\varepsilon^{k+1}} e^{-\pi^2 \varepsilon^2 \|\omega^{(l)}\|^2} \int_0^\varepsilon e^{-\min\{R\|\omega^{(l)}\|, K - \|\omega^{(l)}\|\}} 2\pi \sin(\varepsilon b^{(l)}) db^{(l)} d\omega^{(l)} \right|$$

$$\leq \tau \int_{\|\omega^{(l)}\| \leq K} \frac{2\varepsilon^{k+1}}{2\varepsilon^{k+1}} e^{-\pi^2 \varepsilon^2 \|\omega^{(l)}\|^2} \tau^k d\omega^{(l)} = \tau C_m < C_m.$$ 

For $\xi > 0$, consider the set

$$K^{(l)}_{\xi} \triangleq \left\{ \kappa \in \mathbb{R}^{n \times k} \times \mathbb{R}^n \mid \|\kappa^{(l)}\| \leq K, \|\kappa^{(j)}\| \leq \xi \text{ for } j \neq l \right\}.$$ 

Denote by $B_{\xi}$ the closed ball of radius $\xi$ in $\mathbb{R}^{k+1}$, centered at 0. Now define

$$m^{(l)}_\xi (\kappa^{(1)}, \ldots, \kappa^{(n)}) \triangleq m(\kappa^{(l)}) \frac{1_{K^{(l)}_{\xi}} (\kappa^{(1)}, \ldots, \kappa^{(n)})}{\text{Vol}(B_{\xi})^{n-1}}.$$ 

We will show that for sufficiently small $\xi$ and some constant vector $\nu^{(l)} \in \mathbb{R}^n$,

$$E_Z \left[ \left\| \tilde{\delta}^{(l)} (Z) - \nu^{(l)} - \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \phi(Z; \kappa) m^{(l)}_\xi (\kappa) d\kappa \right\|_2 \right] = O(\varepsilon).$$ 

Note that given $z \in \mathbb{R}^k$,

$$\tilde{\Phi}^{(l)}_{\xi}(z) \triangleq \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \phi(z; \kappa) m^{(l)}_\xi (\kappa) d\kappa = \begin{bmatrix} \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \sigma_g \left( (\omega^{(1)})^T z + b^{(l)} \right) m^{(l)}_\xi (\kappa) d\kappa \\ \vdots \\ \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \sigma_g \left( (\omega^{(n)})^T z + b^{(l)} \right) m^{(l)}_\xi (\kappa) d\kappa \end{bmatrix}.$$
For $j = 1, \ldots, n$, we denote the $j$-th component function of $\tilde{\Phi}_\xi^{(l)}$ by $[\tilde{\Phi}_\xi^{(l)}]_j$. Observe that if we denote $d\kappa^{(-l)} = d\kappa^{(1)} \cdots d\kappa^{(l-1)} d\kappa^{(l)} \cdots d\kappa^{(n)}$, then by our construction of $m$ and $K$,

\[ [\tilde{\Phi}_\xi^{(l)}]_j (Z) = \int_{\mathbb{R}^n} \sigma_{g_j} \left( (\omega^{(l)})^T Z + b^{(l)} \right) m(\kappa^{(l)}) \left( \int_{\mathbb{R}^{(n-1) \times k \times \mathbb{R}^{n-1}}} \frac{1_{\kappa^{(1)}} (\kappa^{(1)}, \ldots, \kappa^{(n)})}{\text{Vol}(B_\xi)^{n-1}} d\kappa^{(-l)} \right) d\kappa^{(l)}, \]

which is $3\varepsilon$-close to $\tilde{\Phi}_\xi^{(l)} - \rho^{(l)}$ within $L^1(\mu, \kappa^{(l)})$, regardless of $\xi$.

Next we bound the remaining components of $\tilde{\Phi}_\xi^{(l)}$. Since $\sigma_g(r)$ is continuous at $r = 0$, we can take $\xi$ so that

\[ |\sigma_g(r) - \sigma_g(0)| < \frac{\varepsilon}{2(n-1)C_m} \quad (11) \]

holds for all $|r| < (1 + R)\xi$. Observe that for $j \neq l$,

\[ [\tilde{\Phi}_\xi^{(l)}]_j (z) = \int_{\mathbb{R}^{(n-1) \times k \times \mathbb{R}^{n-1}}} \sigma_{g_j} \left( (\omega^{(j)})^T z + b^{(j)} \right) \frac{\prod_{m \neq l} 1_{B_\xi} (\kappa^{(m)})}{\text{Vol}(B_\xi)^{1}} \left( \int_{\|\kappa^{(l)}\| \leq K} m(\kappa^{(l)}) d\kappa^{(l)} \right) d\kappa^{(-l)} \]

Define

\[ \rho^{(-l)} \Delta \int_{\|\kappa^{(l)}\| \leq K} \sigma_{g_j} (0) m(\kappa^{(l)}) d\kappa^{(l)} = \frac{1}{\text{Vol}(B_\xi)} \int_{\|\kappa^{(l)}\| \leq K} \sigma_{g} (0) \left( \int_{\|\kappa^{(l)}\| \leq K} m(\kappa^{(l)}) d\kappa^{(l)} \right) d\kappa^{(j)}. \]

Then we have

\[ \left| [\tilde{\Phi}_\xi^{(l)}]_j (z) - \rho^{(-l)} \right| \leq \frac{1}{\text{Vol}(B_\xi)} \int_{\|\kappa^{(l)}\| \leq \xi} \left| \sigma_{g_j} \left( (\omega^{(j)})^T z + b^{(j)} \right) - \sigma_{g_j} (0) \right| \left( \int_{\|\kappa^{(l)}\| \leq K} m(\kappa^{(l)}) d\kappa^{(l)} \right) d\kappa^{(j)} \]

\[ \leq \frac{1}{\text{Vol}(B_\xi)} \int_{\|\kappa^{(l)}\| \leq \xi} C_m \left| \sigma_{g_j} \left( (\omega^{(j)})^T z + b^{(j)} \right) - \sigma_{g_j} (0) \right| d\kappa^{(j)}. \]

Note that the integrand is nonzero only when $\|\kappa^{(l)}\| \leq \xi$, which implies $\|\omega^{(j)}\|, \|b^{(j)}\| \leq \xi$. Therefore, on the event $\|z\| \leq R$, we have $|(\omega^{(j)})^T z + b^{(j)}| \leq \xi (1 + \|z\|) \leq \xi (1 + R)$, so (11) gives

\[ \left| [\tilde{\Phi}_\xi^{(l)}]_j (z) - \rho^{(-l)} \right| \leq C_m \frac{\varepsilon}{2(n-1)C_m} = \frac{\varepsilon}{2(n-1)}. \]

When $\|z\| > R$, the crude bound

\[ \left| [\tilde{\Phi}_\xi^{(l)}]_j (z) - \rho^{(-l)} \right| \leq 2\|\sigma_{g}\|_{\infty} C_m \]

is enough, because $\text{Prob}_{Z} (\|Z\| \geq R) < \frac{\varepsilon}{4(n-1)\|\sigma_{g}\|_{\infty} C_m}$. We have established

\[ \mathbb{E}_{Z} \left[ \left| [\tilde{\Phi}_\xi^{(l)}]_j (Z) - \rho^{(-l)} \right| \right] < \frac{\varepsilon}{n-1} \]

for all $j \neq l$. 
Now, with
\[
\mathbf{v}^{(\epsilon,l)} = \begin{bmatrix} -\rho^{(-l)} \\ \vdots \\ -\rho^{(-l)} \\ -\rho^{(-l)} \\ \vdots \\ -\rho^{(-l)} \end{bmatrix} \quad \text{← on } l\text{-th coordinate,}
\]
we have
\[
\mathbb{E}_Z \left[ \left\| \delta^{(\epsilon,l)}(Z) - \mathbf{v}^{(\epsilon,l)} - \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \phi(Z; \kappa) m^{(l)}_j(\kappa) \, d\kappa \right\|_2 \right] \\
\leq \mathbb{E}_Z \left[ \left\| \delta^{(\epsilon,l)}(Z) - \mathbf{v}^{(\epsilon,l)} - \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \phi(Z; \kappa) m^{(l)}_j(\kappa) \, d\kappa \right\|_1 \right] \\
= \mathbb{E}_Z \left[ \left\| \delta^{(\epsilon)}(Z) - \rho^{(-l)} - \left[ \tilde{\Phi}^{(l)}_j(z) \right] \right\|_1 \right] + \sum_{j \neq l} \mathbb{E}_Z \left[ \left\| \left[ \tilde{\Phi}^{(l)}_j \right] (Z) - \rho^{(-l)} \right\|_1 \right] \\
< 3\varepsilon + (n-1) \frac{\varepsilon}{n-1} = 4\varepsilon.
\]

The space of vector functions \( h = ([h_1], \ldots, [h_n]: \mathbb{R}^k \rightarrow \mathbb{R}^n \) satisfying \( \mathbb{E}_Z \left[ \|[h]_j(Z)\|^2 \right] < \infty \) for each \( j = 1, \ldots, n \) can be identified as the direct sum of \( L^2 \) spaces
\[
\mathcal{H} \triangleq \bigoplus_{j=1}^n L^2(q_Z(z) \, dz).
\]
This is a Hilbert space equipped with the inner product \( \langle g, h \rangle_{\mathcal{H}} = \sum_{j=1}^n \mathbb{E}_Z \left[ [g]_j(Z) [h]_j(Z) \right] = \mathbb{E}_Z \left[ g^\top(Z) h(Z) \right] \). Now let \( p(\kappa) > 0 \) be the density function on \( \mathbb{R}^{n \times k} \times \mathbb{R}^n \) from which we sample \( \kappa \)'s, and define
\[
C^{(l)}_\kappa \triangleq \sup_{\kappa} \frac{m^{(l)}_\kappa(\kappa)}{p(\kappa)},
\]
which is finite because \( m^{(l)}_\kappa \) is bounded and compactly supported, while \( p \) is positive and continuous. For each random \( \kappa_i, i = n+1, \ldots, N_g \), the corresponding realization
\[
h_i := \frac{m^{(l)}_\kappa(\kappa_i)}{p(\kappa_i)} \phi(\cdot; \kappa_i) \in \mathcal{H}
\]
satisfies \( \|h_i\|_{\mathcal{H}} \leq \sqrt{n} \|\sigma_g\|_{\infty} C^{(l)}_{\kappa} \). Hence, as in the \( n = 1 \) case,
\[
\left\| \frac{1}{N_g - n} \sum_{i=n+1}^{N_g} h_i - \int_{\mathbb{R}^{n \times k} \times \mathbb{R}^n} \phi(Z; \kappa) m^{(l)}_\kappa(\kappa) \, d\kappa \right\|_{\mathcal{H}} \leq \frac{\sqrt{n} \|\sigma_g\|_{\infty} C^{(l)}_{\kappa}}{\sqrt{N_g - n}} \left( 1 + \sqrt{2 \log \frac{1}{\zeta}} \right)
\]
with probability \( \geq 1 - \zeta \) over \( \kappa_{n+1}, \ldots, \kappa_{N_g} \). Let \( \theta^{(\epsilon,l)}_i = \frac{m^{(l)}_\kappa(\kappa_i)}{(N_g - n)p(\kappa_i)} \) for \( i = n+1, \ldots, N_g \). Take \( \kappa_1 = (0_{n \times k}, b_1), \ldots, \kappa_n = (0_{n \times k}, b_n) \), where \( b_1, \ldots, b_n \in \mathbb{R}^k \), in a way that the constant vectors \( \sigma_g(b_i) \) are linearly independent. Then there exist \( \theta_{\delta}^{(\epsilon,l)}(b_i) = \sum_{i=1}^{n} \theta_i^{(\epsilon,l)}(b_i) \cdot \sigma_g(b_i) = \mathbf{v}^{(\epsilon,l)} \).
Given \( f: \mathbb{R}^k \to \mathbb{R}^n \), let \( M = \sup_{z \in \mathbb{R}^k} \| f(z) \|_2 \). Chaining all the approximation steps, we have

\[
\begin{align*}
\mathbb{E}_Z \left[ \left( \delta^{(l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(e,l)} \phi(Z; \kappa_i) \right)^\top f(Z) \right] & \leq \mathbb{E}_Z \left[ \left( \delta^{(l)}(Z) - \tilde{\delta}^{(e,l)}(Z) \right)^\top f(Z) \right] + \mathbb{E}_Z \left[ \left( \tilde{\delta}^{(e,l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(e,l)} \phi(Z; \kappa_i) \right)^\top f(Z) \right] \\
& \leq \mathbb{E}_Z \left[ \left( \delta(Z) - \tilde{\delta}(Z) \right)^\top f(Z) \right] + M \mathbb{E}_Z \left[ \left\| \delta^{(e,l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(e,l)} \phi(Z; \kappa_i) \right\|_2 \right] \\
& \leq C_\delta \varepsilon \sup_{z \in \mathbb{R}^k} \left( \| f(z) \| + \| \nabla f(z) \| \right) + M \mathbb{E}_Z \left[ \left\| \tilde{\delta}^{(e,l)}(Z) - \int_{\mathbb{R}^k \times \mathbb{R}^n} \phi(Z; \kappa) m_\xi^{(l)}(\kappa) d\kappa - \sum_{i=1}^{N_g} \theta_i^{(e,l)} \phi(Z; \kappa_i) \right\|_2 \right] \\
& \leq \left( (C_\delta + 4) \varepsilon + \frac{\sqrt{\dim} \| \sigma_g \|_\infty C_\nu^{(l)}}{\sqrt{N_g - n}} \left( 1 + \sqrt{2 \log \frac{1}{\zeta}} \right) \right) \sup_{z \in \mathbb{R}^k} \left( \| f(z) \|_2 + \| Df(z) \| \right)
\end{align*}
\]

with probability \( \geq 1 - \zeta \). Clearly, with sufficiently large \( N_g \), the last term is \( \mathcal{O}(\varepsilon \sup_{z \in \mathbb{R}^k} (\| f(z) \|_2 + \| Df(z) \|)) \).

### B. Experimental details and additional experimental results

#### B.1. Gaussian mixture sample generation

We now provide details of the experiments for Figure 4. The code is available at [https://github.com/sehyunkwon/Infinite-WGAN](https://github.com/sehyunkwon/Infinite-WGAN). The true and latent distributions are 2-dimensional, i.e., \( n = 2 \) and \( k = 2 \). The true distribution \( P_X \) is a mixture of 8 Gaussians with equal weights, where the means are \( \left( \sqrt{2} \cos \frac{m\pi}{4}, \sqrt{2} \sin \frac{m\pi}{4} \right) \) for \( m = 0, 1, \ldots, 7 \), and the covariance matrices are \( \begin{pmatrix} 0.1^2 & 0 \\ 0 & 0.1^2 \end{pmatrix} \). The generator feature functions are of the form \( \phi_i(x) = \sigma_g(\kappa_w x + \kappa_b) \) as described in (AG) for \( 1 \leq i \leq N_g = 5000 \), where the activation function \( \sigma_g \) is tanh. Weights \( \kappa_w \) and \( \kappa_b \) for generator feature functions are randomly sampled (IID) from the Gaussian distribution with zero mean and variance \( 10^2 \). As required in Lemma 7, we create constant hidden units by replacing two sets of \( (\kappa_w, \kappa_b) \) with \( (0, 2, (1, 0)) \) and \( (0, 2, (0, 1)) \). The discriminator feature functions are of form \( \psi_j(x) = \sigma(a_j^T x + b_j) \) as described in (AD) for \( 1 \leq j \leq N_d = 1000 \) where the activation function \( \sigma \) is tanh. We generate \( a_j \) and \( b_j \) independently according to the following procedure:

- Pick \( x \)-intercept \( \tilde{a} \) and \( y \)-intercept \( \tilde{b} \) from \(-4 \) to 4 uniformly randomly.
- Then, \( \frac{\tilde{a}}{\tilde{b}} + \frac{\tilde{b}}{\tilde{a}} = 1 \) is the line with those intercepts.
- Pick \( c \) uniformly randomly from 1 to 10, then set \( a_j = (c/\tilde{a}, c/\tilde{b}) \) and \( b_j = -c \).

The generator stepsize starts at \( \alpha = 10^{-5} \) and decays by a factor of 0.9 at every epoch. The networks are trained for 25 epochs with \( (X\text{-sample}) \) batch size 5000. At each iteration, 5000 latent vectors \( (Z\text{-samples}) \) are sampled IID from the standard Gaussian distribution for the stochastic gradient ascent step, and another 5000 latent vectors are sampled IID from the standard Gaussian distribution for the stochastic gradient descent step. The generator parameter \( \theta \) is randomly initialized (IID) with the Gaussian distribution with zero mean and variance \( 5 \times 10^{-3} \). The discriminator parameter \( \gamma \) is randomly initialized (IID) with the standard normal distribution. We visualize the generated distribution using the kernel density estimation (KDE) plot.

We also perform additional experiments under distinct settings. The first additional experiment considers the true distribution \( P_X \) that is a mixture of 9 Gaussians with equal weights. The means are \( (m_1, m_2) \) for \( m_1 = -1, 0, 1 \) and \( m_2 = -1, 0, 1 \), and the covariance matrices are the same as before. We use the initial stepsize \( \alpha = 5 \times 10^{-6} \) for the generator and
$N_g = 10,000$ for generator feature functions. The generator parameter $\theta$ is randomly initialized (IID) with the Gaussian distribution with zero mean and variance $3 \times 10^{-3}$. Discriminator feature functions are generated in the same manner. Figure 4(a) shows the true distribution, and Figure 4(b) shows the generated samples. The second additional experiment considers the true distribution $P_X$, which is a spiral-shaped mixture of 20 Gaussians with equal weights. The means are $(m_20 \cos \frac{2m_20 \pi}{20}, m_20 \sin \frac{2m_20 \pi}{20})$ for $m = 0, 1, \ldots, 19$, and the covariance matrices are the same as before. We use the initial stepsize $\alpha = 10^{-6}$ for the generator and $N_g = 10,000$ for generator feature functions. The generator parameter $\theta$ is randomly initialized (IID) with the Gaussian distribution with zero mean and variance $3 \times 10^{-3}$. For the discriminator, feature function weights are generated by sampling $x$-intercept $\tilde{a}$ and $y$-intercept $\tilde{b}$ from $-2$ to $2$ uniformly randomly. Figure 4(d) shows the true distribution, and Figure 4(e) shows the generated samples. In both cases, the generators closely mimic the true distributions and loss functions converge to zero.

**B.2. Loss landscape**

In this section, we describe the experiments for Figure 3, which visualizes the loss landscape of $J(\theta)$ for the cases $N_g = 2$ and $N_g = 10$. We also provide additional experiments for $N_g = 3, 5,$ and $100$. In the $N_g = 2$ case, the landscape is highly non-convex and displays at least three non-global local minima. We observe that in Figures 5 and 6, the landscapes become better behaved, although still non-convex, as $N_g$ increases.

When $N_g > 2$, the parameter space is projected down to a 2D plane spanned by two random directions, as recommended by Li et al. (2018b). The true and latent distributions are 2-dimensional, i.e., $n = 2$ and $k = 2$. The true distribution $P_X$ is a mixture of 2 Gaussians with equal weights, where the means are $(m_1, m_2)$ for $m_1 = 0$ and $m_2 = \pm 2$, and the
covariance matrices are \(
\begin{pmatrix}
\sqrt{0.5^2} & 0 \\
0 & \sqrt{0.5^2}
\end{pmatrix}
\). The latent distribution is the standard Gaussian distribution. The generator feature functions are of the form \(\phi_i(x) = \sigma_g(\kappa_w z + \kappa_b)\) as described in (AG) for \(1 \leq i \leq N_g = 2, 3, 5, 10,\) and \(100\), where the activation function \(\sigma_g\) is \(\tanh\). Weights \(\kappa_w\) are randomly sampled (IID) from an isotropic Gaussian and then multiplied by a scalar factor, sampled independently from the standard normal distribution. Weights \(\kappa_b\) are randomly sampled (IID) from the Gaussian distribution with zero mean and variance \(3 \times 10^{-1}\). The discriminator feature functions are of the form \(\psi_j(x) = \sigma(a_j^T x + b_j)\) as described in (AD) for \(1 \leq j \leq N_d = 8\), where the activation function \(\sigma\) is \(\tanh\).
Figure 5. Loss landscapes of $J(\theta)$ for $N_g = 2$, 3, 5, 10, and 100.

Figure 6. Corresponding contour plots of the landscapes of Figure 5.