A. Omitted proofs

A.1. Proof of Lemma 1

Theorem 9 ((Hornik, 1991, Theorem 1)). *Let $\sigma\colon \mathbb{R} \to \mathbb{R}$ 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 $\{(\theta_i, a_i, b_i) \in \mathbb{R} \times \mathbb{R}^k \times \mathbb{R}\}_{i=1}^N$ such that*

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

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

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

$$h_l(z) = \sum_{i=1}^{N_l} \theta_i^{(l)} \sigma_{\mathrm{g}}((a_i^{(l)})^\intercal z - b_i^{(l)})$$

satisfies

$$\int_{\mathbb{R}^k} |h_l(z) - f_l(z)|\, q_Z(z) dz < \frac{\varepsilon}{2n}, \tag{3}$$

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

$$A_i^{(l)} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ (a_i^{(l)})^\intercal \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow l\text{-th row}, \qquad \mathbf{b}_i^{(l),r} = \begin{bmatrix} r \\ \vdots \\ r \\ b_i^{(l)} \\ r \\ \vdots \\ r \end{bmatrix}, \qquad e_{-l} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \leftarrow \text{on } l\text{-th coordinates}$$

and

$$\tilde{f}^{(l),r}(z) = -\left(\sum_{i=1}^{N_l} \theta_i^{(l)} \right) \ell_g e_{-l} + \sum_{i=1}^{N_l} \theta_i^{(l)} \sigma_{\mathrm{g}}(A_i^{(l)} z - \mathbf{b}_i^{(l),r}).$$

Then, for each $l = 1, \ldots, n$, we have $\tilde{f}_j^{(l),r} = \sum_{i=1}^{N_l} \theta_i^{(l)} (\sigma_g(-r) - \ell_g) \to 0$ as $r \to \infty$ if $j \neq l$, while $\tilde{f}_l^{(l),r} = h_l(z)$. Because σ_g 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 $r_{\mathrm{big}} > 0$ such that

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

and we conclude with (3) that

$$\int_{\mathbb{R}^k} \left\| f(z) - \sum_{l=1}^n \tilde{f}^{(l),r_{\mathrm{big}}}(z) \right\|_1 q_Z(z) dz < \varepsilon.$$

Note that

$$\sum_{l=1}^{n} \tilde{f}^{(l),r_{\text{big}}}(z) \in \overline{\text{span}}(\mathcal{G}).$$

Therefore, using the bound $\|\cdot\|_2 \le \|\cdot\|_1$, we get

$$\int_{\mathbb{R}^k} \left\| f(z) - \sum_{l=1}^{n} \tilde{f}^{(l),r_{\text{big}}}(z) \right\|_2 q_Z(z)dz < \varepsilon.$$

□

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 $\mathbb{E}_Z\left[\|h(Z)\|_2\right] < \infty$. Therefore, for any $\varepsilon > 0$, there exists θ_ε such that $\mathbb{E}_Z\left[\|g_{\theta_\varepsilon}(Z) - h(Z)\|_2\right] < \varepsilon$. Observe that

$$\begin{aligned}
\mathbb{E}_Z\left[g^{\intercal}_{\theta_\varepsilon}(Z)h(Z)\right] &= \int_{\mathbb{R}^k}\int_{\mathbb{R}^p} h^{\intercal}(z)\phi(z;\kappa)\,d\theta_\varepsilon(\kappa)\,q_Z(z)\,dz \\
&= \int_{\mathbb{R}^p}\int_{\mathbb{R}^k} h^{\intercal}(z)\phi(z;\kappa)q_Z(z)\,dz\,d\theta_\varepsilon(\kappa) \\
&= \int \mathbb{E}_Z\left[h^{\intercal}(Z)\phi(Z;\kappa)\right] d\theta_\varepsilon(\kappa) = 0.
\end{aligned}$$

Here the change in the order of integration is valid because $\phi(z;\kappa) = \sigma_{\text{g}}(\kappa_w z + \kappa_b) \le \|\sigma_{\text{g}}\|_\infty$ and the total mass of θ_ε is finite, so that

$$\int_{\mathbb{R}^k}\int_{\mathbb{R}^p} \|h^{\intercal}(z)\phi(z;\kappa)\|_2 \; d\theta_\varepsilon(\kappa)\,q_Z(z)\,dz \le n\|h\|_\infty \|\sigma_{\text{g}}\|_\infty \theta_\varepsilon\left(\mathbb{R}^p\right) < \infty.$$

To clarify, the $\|\cdot\|_\infty$ for $\|\sigma_{\text{g}}\|_\infty$ is the standard supremum norm for L^∞ spaces while $\|h\|_\infty = \max_{1\le i\le n}\|h_i\|_\infty$ where $h_i(z)$ is the i-th coordinate of $h(z) \in \mathbb{R}^n$. Finally, we have

$$\begin{aligned}
\mathbb{E}_Z\left[\|h(Z)\|_2^2\right] = \mathbb{E}_Z\left[h^{\intercal}(Z)\left(h(Z) - g_{\theta_\varepsilon}(Z)\right)\right] &\le \|h\|_\infty \mathbb{E}_Z\left[\|h(Z) - g_{\theta_\varepsilon}(Z)\|_1\right] \\
&\le \|h\|_\infty \mathbb{E}_Z\left[\sqrt{n}\,\|h(Z) - g_{\theta_\varepsilon}(Z)\|_2\right] < \varepsilon\sqrt{n}\|h\|_\infty.
\end{aligned}$$

To clarify, $\|h(Z) - g_{\theta_\varepsilon}(Z)\|_1$ denotes the ℓ^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 $z \in \mathbb{R}^k$. □

A.3. Proof of Lemma 5

Theorem 10 ((Sussmann, 1992, Lemma 1)). *Let $\sigma = \tanh$. Assume*

$$C_0 + \sum_{j=1}^{N} \eta_j \sigma(a_j^{\intercal}x + b_j) = C$$

for all $x \in \mathbb{R}^n$, where $\eta_j \neq 0$ and $a_j \neq 0$ for $1 \le j \le N$. 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 $\sigma = \tanh$. With probability 1, the condition of Theorem 10 holds, and

$$F(x) \triangleq \sum_{j=1}^{N_d} \eta_j \psi_j(x)$$

with $\eta \neq 0$ is not constant. Since $\sigma\colon \mathbb{R} \to \mathbb{R}$ is analytic on $\mathbb{R}$, it has a power series expansion

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

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

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

for $x \in B$, and

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

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

$$\begin{aligned}
F(x_u(t)) &= \sum_{j=1}^{N_d} \eta_j \sum_{\nu=0}^{\infty} s_\nu (t\alpha_j + \beta_j)^\nu \\
&= \sum_{j=1}^{N_d} \eta_j \sum_{\nu=0}^{\infty} s_\nu \sum_{m=0}^{\nu} \binom{\nu}{m} (\alpha_j t)^m \beta_j^{\nu-m} \\
&= \sum_{m=0}^{\infty} \left(\sum_{j=1}^{N_d} \sum_{\nu\geq m} \eta_j s_\nu \binom{\nu}{m} \alpha_j^m \beta_j^{\nu-m} \right) t^m \\
&\triangleq F_0 + \sum_{m=1}^{\infty} F_m t^m
\end{aligned}$$

is constant within $t \in (-\varepsilon, \varepsilon)$ for some $\varepsilon > 0$. (Order of summations can be freely interchanged because power series for σ 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. □

A.4. Proof of Lemma 7

Recall that we defined

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

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

Lemma 11. *Assume (AL). There exists a constant C_δ depending only on k but not on $\varepsilon > 0$ such that*

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

for all differentiable $f\colon \mathbb{R}^k \to \mathbb{R}$ 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 ℓ^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 $q_Z(z)$. Then for any $z \in \mathbb{R}^k$,

$$|f(z)q_Z(z) - f(0)q_Z(0)| \le |f(z)||q_Z(z) - q_z(0)| + |f(z) - f(0)|q_Z(0) \le ML_Z\|z\| + L_f\|z\|q_Z(0).$$

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

$$\begin{aligned}\int_{\mathbb{R}^k} |f(z)q_Z(z) - f(0)q_Z(0)|\,\tilde{\delta}^\varepsilon(z)\,dz &\le \int_{\mathbb{R}^k} (ML_Z + L_f q_Z(0))\|z\|\tilde{\delta}^\varepsilon(z)\,dz \\ &= \int_{\mathbb{R}^k} (ML_Z + L_f q_Z(0))\frac{\pi^{-k/2}}{\varepsilon^k}\|z\|e^{-\|z/\varepsilon\|_2^2}\,dz.\end{aligned}$$

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

$$\begin{aligned}(ML_Z + L_f q_Z(0))\pi^{-k/2}\varepsilon\int_{\mathbb{R}^k}\|z\|e^{-\|z\|_2^2}\,dz &\le \max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k}\|z\|e^{-\|z\|_2^2}\,dz\right)\varepsilon(M + L_f) \\ &\le 2\max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k}\|z\|e^{-\|z\|_2^2}\,dz\right)\varepsilon\sup_{z\in\mathbb{R}^k}\left(|f(z)| + \|Df(z)\|\right),\end{aligned}$$

which shows that

$$\begin{aligned}\left|\mathbb{E}_Z\left[\left(\tilde{\delta}^\varepsilon(Z) - \delta(Z)\right)f(Z)\right]\right| &\le \int_{\mathbb{R}^k} |f(z)q_Z(z) - f(0)q_Z(0)|\tilde{\delta}^\varepsilon(z)\,dz \\ &\le C_\delta\,\varepsilon\sup_{z\in\mathbb{R}^k}\left(|f(z)| + \|Df(z)\|\right)\end{aligned}$$

where

$$C_\delta = 2\max\{L_Z, q_Z(0)\}\pi^{-k/2}\left(\int_{\mathbb{R}^k}\|z\|e^{-\|z\|_2^2}\,dz\right).$$

□

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. Approximate $\delta(z)$ with $\tilde{\delta}^\varepsilon(z)$ in the sense of Lemma 11.

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

$$\tilde{\delta}^\varepsilon(z) - \tilde{\delta}^\varepsilon(0) = \mathrm{Re}\int\left(e^{2\pi i\omega^\intercal z} - 1\right)\mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\,d\omega = \int\left(\cos\left(2\pi\omega^\intercal z\right) - 1\right)\mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\,d\omega \tag{4}$$

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

$$\int_{\|z\|>R} q_Z(z)\,dz < \frac{\pi^{k/2}\varepsilon^{k+1}}{2}.$$

Following (Telgarsky, 2020, Section 4.2), for $\|z\| \le R$, the cosine term in (4) can be rewritten as

$$\begin{aligned}
&\cos\left(2\pi\omega^\intercal z\right) - 1 \\
&= \int_0^{\omega^\intercal z} -2\pi \sin(2\pi b)\,db \\
&= \int_0^{R\|\omega\|} -2\pi\,\mathbf{1}_{\{\omega^\intercal z - b \ge 0\}}(z)\,\sin(2\pi b)\,db + \int_{-R\|\omega\|}^{0} 2\pi\,\mathbf{1}_{\{\omega^\intercal z - b \le 0\}}(z)\,\sin(2\pi b)\,db.
\end{aligned} \tag{5}$$

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

$$\mathbf{1}_{\{r\ge 0\}}(r) = \lim_{\tau\downarrow 0}\frac{1}{u_g - \ell_g}\left(\sigma_g\left(\frac{r}{\tau}\right) - \ell_g\right)$$

for $r \neq 0$. Hence we can approximate the step function terms in (5) using σ_g:

$$\int_0^{R\|\omega\|} \lim_{\tau\downarrow 0} -\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{\omega^\intercal z - b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,db + \int_{-R\|\omega\|}^{0} \lim_{\tau\downarrow 0}\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{-\omega^\intercal z + b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,db. \tag{6}$$

Plugging (6) into (4), we obtain

$$\begin{aligned}
&\tilde\delta^\varepsilon(z) - \tilde\delta^\varepsilon(0) \\
&= \int\int_0^{R\|\omega\|} \lim_{\tau\downarrow 0} -\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{\omega^\intercal z - b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,\mathcal{F}[\tilde\delta^\varepsilon](\omega)\,db\,d\omega \\
&\quad + \int\int_{-R\|\omega\|}^{0} \lim_{\tau\downarrow 0}\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{-\omega^\intercal z + b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,\mathcal{F}[\tilde\delta^\varepsilon](\omega)\,db\,d\omega
\end{aligned} \tag{7}$$

for $\|z\| \le R$.

Observe that because σ_g is bounded and by Lemma 12, for any $\tau > 0$ and $z \in \mathbb{R}^k$,

$$\int\int_0^{R\|\omega\|} \left|\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{\omega^\intercal z - b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,\mathcal{F}[\tilde\delta^\varepsilon](\omega)\right| db\,d\omega \le \int \frac{2\pi\left(\|\sigma_g\|_\infty + \ell_g\right)}{u_g-\ell_g} R\|\omega\|\,\mathcal{F}[\tilde\delta^\varepsilon]\,d\omega < \infty.$$

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 $\tilde\delta^\varepsilon(z)$ as

$$\begin{aligned}
\tilde\delta^\varepsilon(z) &= \tilde\delta^\varepsilon(0) + \lim_{\tau\downarrow 0}\int\int_0^{R\|\omega\|} -\frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{\omega^\intercal z - b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,\mathcal{F}[\tilde\delta^\varepsilon](\omega)\,db\,d\omega \\
&\quad + \lim_{\tau\downarrow 0}\int\int_{-R\|\omega\|}^{0} \frac{2\pi}{u_g-\ell_g}\left(\sigma_g\left(\frac{-\omega^\intercal z + b}{\tau}\right) - \ell_g\right)\sin(2\pi b)\,\mathcal{F}[\tilde\delta^\varepsilon](\omega)\,db\,d\omega \\
&= \theta_1^\varepsilon\phi(z;\kappa_1) + \lim_{\tau\downarrow 0}\int\int_{-R\|\omega\|}^{0} -\frac{2\pi\tau^{k+1}}{u_g-\ell_g}\sigma_g\left(\omega^\intercal z + b\right)\sin(-2\pi\tau b)\,\mathcal{F}[\tilde\delta^\varepsilon](\tau\omega)\,db\,d\omega \\
&\quad + \lim_{\tau\downarrow 0}\int\int_{-R\|\omega\|}^{0} \frac{2\pi\tau^{k+1}}{u_g-\ell_g}\sigma_g\left(\omega^\intercal z + b\right)\sin(2\pi\tau b)\,\mathcal{F}[\tilde\delta^\varepsilon](-\tau\omega)\,db\,d\omega \\
&= \theta_1^\varepsilon\phi(z;\kappa_1) + \lim_{\tau\downarrow 0}\int_{\mathbb{R}^k\times\mathbb{R}} \phi(z;\kappa)m_\tau(\kappa)\,d\kappa
\end{aligned}$$

for $\|z\| \le R$. We specify the notations that were newly introduced. First, we denoted $\kappa = (\omega, b)$, so that $\phi(z;\kappa) = \sigma_g(\omega^\intercal z + b)$ (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

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

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

$$\begin{aligned} m_\tau(\kappa) &= \frac{2\pi\tau^{k+1}}{u_g - \ell_g} \left(-\sin(-2\pi\tau b)\, \mathcal{F}[\tilde{\delta}^\varepsilon](\tau\omega)\, \mathbf{1}_{\{-R\|\omega\| \le b \le 0\}}(\kappa) + \sin(2\pi\tau b)\, \mathcal{F}[\tilde{\delta}^\varepsilon](-\tau\omega)\, \mathbf{1}_{\{-R\|\omega\| \le b \le 0\}}(\kappa) \right) \\ &= \frac{4\pi\tau^{k+1}}{u_g - \ell_g} e^{-\pi^2\varepsilon^2\tau^2\|\omega\|^2} \sin(2\pi\tau b)\, \mathbf{1}_{\{-R\|\omega\| \le b \le 0\}}(\kappa), \end{aligned} \tag{8}$$

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

$$\begin{aligned} &\cos(2\pi\omega^\intercal z) - 1 - \int_0^{R\|\omega\|} -2\pi\, \mathbf{1}_{\{\omega^\intercal z - b \ge 0\}}(z)\, \sin(2\pi b)\, db + \int_{-R\|\omega\|}^0 2\pi\, \mathbf{1}_{\{\omega^\intercal z - b \le 0\}}(z)\, \sin(2\pi b)\, db \\ &= \left(\cos(2\pi\omega^\intercal z) - \cos(2\pi R\|\omega\|)\right) \mathbf{1}_{\{|\omega^\intercal z| > R\|\omega\|\}}(\omega), \end{aligned}$$

and thus

$$\tilde{\delta}^\varepsilon(z) - \theta_1^\varepsilon \phi(z;\kappa_1) - \lim_{\tau \downarrow 0} \int_{\mathbb{R}^k \times \mathbb{R}} \phi(z;\kappa) m_\tau(\kappa)\, d\kappa = \int_{\{\omega \,|\, |\omega^\intercal z| > R\|\omega\|\}} \left(\cos(2\pi\omega^\intercal z) - \cos(2\pi R\|\omega\|)\right) \mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\, d\omega \tag{9}$$

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

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

Therefore, the family

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

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

$$\begin{aligned} &\lim_{\tau \downarrow 0} \mathbb{E}_Z \left[\left| \tilde{\delta}^\varepsilon(Z) - \theta_1^\varepsilon \phi(Z;\kappa_1) - \int_{\mathbb{R}^k \times \mathbb{R}} \phi(Z;\kappa) m_\tau(\kappa)\, d\kappa \right| \right] \\ &= \mathbb{E}_Z \left[\left| \int_{\{\omega \,|\, |\omega^\intercal Z| > R\|\omega\|\}} \left(\cos(2\pi\omega^\intercal Z) - \cos(2\pi R\|\omega\|)\right) \mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\, d\omega \right| \right] \\ &\le \mathbb{E}_Z \left[\int_{\{\omega \,|\, |\omega^\intercal Z| > R\|\omega\|\}} 2\mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\, d\omega \right] \\ &\le \mathbb{E}_Z \left[\mathbf{1}_{\{\|z\| > R\}}(Z) \int_{\mathbb{R}^k} 2\mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\, d\omega \right] \\ &= \left(\int_{\{\|z\| > R\}} q_Z(z)\, dz \right) \left(\int_{\mathbb{R}^k} 2\mathcal{F}[\tilde{\delta}^\varepsilon](\omega)\, d\omega \right) < \frac{\pi^{k/2}\varepsilon^{k+1}}{2} \frac{2}{\pi^{k/2}\varepsilon^k} = \varepsilon. \end{aligned}$$

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 $\mathbb{E}_Z\left[\left|\tilde{\delta}^\varepsilon(Z) - \theta_1^\varepsilon\phi(Z;\kappa_1) - \int_{\mathbb{R}^k\times\mathbb{R}}\phi(Z;\kappa)m_\tau(\kappa)\,d\kappa\right|\right] < 2\varepsilon$. Because σ_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

$$\begin{aligned}
&\mathbb{E}_Z\left[\left|\tilde{\delta}^\varepsilon(Z)f(Z) - \theta_1^\varepsilon\phi(Z;\kappa_1)f(Z) - \int_{\|\kappa\|\le K}\phi(Z;\kappa)f(Z)m_\tau(\kappa)\,d\kappa\right|\right] \\
&\le \mathbb{E}_Z\left[\left|\tilde{\delta}^\varepsilon(Z)f(Z) - \theta_1^\varepsilon\phi(Z;\kappa_1)f(Z) - \int_{\mathbb{R}^k\times\mathbb{R}}\phi(Z;\kappa)f(Z)m_\tau(\kappa)\,d\kappa\right|\right] \\
&\quad + \mathbb{E}_Z\left[\int_{\|\kappa\|>K}|\phi(Z;\kappa)f(Z)m_\tau(\kappa)|\;d\kappa\right] \\
&\le 2\varepsilon\|f\|_\infty + \|\sigma_g\|_\infty\|f\|_\infty\mathbb{E}_Z\left[\int_{\|\kappa\|>K}|m_\tau(\kappa)|\,d\kappa\right] \le 3\varepsilon\|f\|_\infty.
\end{aligned}$$

Step 4. Approximate the integral over a finite ball by a finite linear combination of random functions in $\mathcal{G}$.

Define

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

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

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

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

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

with high probability over $\kappa_2,\ldots,\kappa_{N_g}$. (The indexing begins with $i = 2$ because κ_1 is reserved for the constant function.) When we draw each κ_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 $\|h_i\|_\infty \le \|\sigma_g\|_\infty C_K$ for all $i = 2,\ldots,N_g$, which implies $\|h_i\|_{\mathcal{H}} \le \|\sigma_g\|_\infty C_K$. 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

$$\mathbb{E}_{\kappa\sim p(\kappa)}\left[\frac{m_{\tau,K}(\kappa)}{p(\kappa)}\phi(\cdot;\kappa)\right] = \int\phi(\cdot;\kappa)\frac{m_{\tau,K}(\kappa)}{p(\kappa)}p(\kappa)\,d\kappa = \int_{\|\kappa\|\le K}\phi(\cdot;\kappa)\,m_\tau(\kappa)\,d\kappa.$$

Therefore, applying the McDiarmid-type bound from (Rahimi & Recht, 2007, Lemma 4), we get

$$\left\| \frac{1}{N_g - 1} \sum_{i=2}^{N_g} h_i - \mathbb{E}_{\kappa \sim p(\kappa)} \left[\frac{m_{\tau,K}(\kappa)}{p(\kappa)} \phi(\cdot;\kappa) \right] \right\|_{\mathcal{H}} = \left\| \sum_{i=2}^{N_g} \frac{m_{\tau,K}(\kappa_i)}{(N_g-1)p(\kappa_i)} \phi(\cdot;\kappa_i) - \int_{\|\kappa\| \le K} \phi(\cdot;\kappa)\, m_\tau(\kappa)\, d\kappa \right\|_{\mathcal{H}} \quad (10)$$
$$\le \frac{\|\sigma_g\|_\infty C_K}{\sqrt{N_g - 1}} \left(1 + \sqrt{2 \log \frac{1}{\zeta}} \right),$$

with probability at least $1 - \zeta$ over $\kappa_2, \ldots, \kappa_{N_g}$. Fix N_g large enough so that the right hand side of (10) is less than ε, and let $\theta_i^\varepsilon = \frac{m_{\tau,K}(\kappa_i)}{(N_g-1)p(\kappa_i)}$ for $i = 2, \ldots, N_g$. Then, using Jensen's inequality, we obtain

$$\varepsilon > \left\| \sum_{i=2}^{N_g} \theta_i^\varepsilon \phi(\cdot;\kappa_i) - \int_{\|\kappa\| \le K} \phi(\cdot;\kappa)\, m_\tau(\kappa)\, d\kappa \right\|_{\mathcal{H}} = \left(\mathbb{E}_Z \left[\left| \sum_{i=2}^{N_g} \theta_i^\varepsilon \phi(Z;\kappa_i) - \int_{\|\kappa\| \le K} \phi(Z;\kappa)\, m_\tau(\kappa)\, d\kappa \right|^2 \right] \right)^{1/2}$$
$$\ge \mathbb{E}_Z \left[\left| \sum_{i=2}^{N_g} \theta_i^\varepsilon \phi(Z;\kappa_i) - \int_{\|\kappa\| \le K} \phi(Z;\kappa)\, m_\tau(\kappa)\, d\kappa \right| \right]$$

with probability $\ge 1 - \zeta$.

Step 5. Combine Steps 1 through 4.

Let $\kappa_1, \ldots, \kappa_{N_g}$ be as above, and $\phi_i(z) = \phi(z;\kappa_i)$. For any continuously differentiable function $f : \mathbb{R}^k \to \mathbb{R}$ such that $\sup_{z \in \mathbb{R}^k} (|f(z)| + \|\nabla f(z)\|) < \infty$, with probability at least $1 - \zeta$, we have

$$\begin{aligned}
&\left| \mathbb{E}_Z \left[\left(\delta(Z) - \sum_{i=1}^{N_g} \theta_i^\varepsilon \phi_i(Z) \right) f(Z) \right] \right| \\
&\le \mathbb{E}_Z \left[\left| \left(\delta(Z) - \tilde{\delta}^\varepsilon(Z) \right) f(Z) \right| \right] + \mathbb{E}_Z \left[\left| \tilde{\delta}^\varepsilon(Z) f(Z) - \theta_1^\varepsilon \phi_1(Z) f(Z) - \int_{\|\kappa\| \le K} \phi(Z;\kappa) f(Z)\, m_\tau(\kappa)\, d\kappa \right| \right] \\
&\quad + \mathbb{E}_Z \left[\left| \theta_1^\varepsilon \phi_1(Z) f(Z) + \int_{\|\kappa\| \le K} \phi(Z;\kappa) f(Z)\, m_\tau(\kappa)\, d\kappa - \sum_{i=1}^{N_g} \theta_i^\varepsilon \phi_i(Z) f(Z) \right| \right] \\
&\le C_\delta\, \varepsilon \sup_{z \in \mathbb{R}^k} (|f(z)| + \|\nabla f(z)\|) + 3\varepsilon \|f\|_\infty + \|f\|_\infty \mathbb{E}_Z \left[\left| \int_{\|\kappa\| \le K} \phi(Z;\kappa)\, m_\tau(\kappa)\, d\kappa - \sum_{i=2}^{N_g} \theta_i^\varepsilon \phi_i(Z) \right| \right] \\
&\le (C_\delta + 4)\, \varepsilon \sup_{z \in \mathbb{R}^k} (\|f(z)\| + \|\nabla f(z)\|),
\end{aligned}$$

where C_δ is the constant (not depending on ε) defined in Lemma 11. This completes the proof for the case $n = 1$.

Proof for the general case $n > 1$.

The crux of the general case is that κ cannot be sampled coordinate-wisely, but we must keep only one coordinate active, while suppressing the others. To achieve this, we simply accept κ'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

$$\kappa = \begin{bmatrix} \kappa^{(1)} \\ \vdots \\ \kappa^{(n)} \end{bmatrix} = \begin{bmatrix} \left(\omega^{(1)}\right)^\intercal & b^{(1)} \\ \vdots & \vdots \\ \left(\omega^{(n)}\right)^\intercal & b^{(n)} \end{bmatrix},$$

where $\omega^{(j)} \in \mathbb{R}^k$, $b^{(j)} \in \mathbb{R}$ and $\kappa^{(j)} = \left[\left(\omega^{(j)}\right)^\intercal, b^{(j)} \right]$ for $j = 1, \ldots, n$.

Fix $1 \le l \le n$ and $\varepsilon > 0$. Define

$$\tilde{\delta}^{(\varepsilon,l)}(z) \triangleq \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \tilde{\delta}^{\varepsilon}(z) \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow \text{on } l\text{-th coordinate.}$$

Let

$$C_m \triangleq \frac{4}{u_g - \ell_g} \int_{\mathbb{R}^k} e^{-\pi^2 \varepsilon^2 \|\omega\|^2} \, d\omega,$$

which is a constant depending only on ε. Take a large $R > 0$ satisfying

$$\int_{\|z\|>R} q_Z(z) \, dz < \min\left\{ \frac{\pi^{k/2} \varepsilon^{k+1}}{2}, \frac{\varepsilon}{4(n-1)\|\sigma_g\|_\infty C_m} \right\}.$$

From Steps 2 and 3 in the case $n = 1$, we can find a density function $m = m_\tau$ of the form

$$m_\tau(\kappa^{(l)}) = \frac{4\pi\tau^{k+1}}{u_g - \ell_g} e^{-\pi^2\varepsilon^2\tau^2\|\omega^{(l)}\|^2} \sin(2\pi\tau b^{(l)}) \, \mathbf{1}_{\{-R\|\omega^{(l)}\| \le b^{(l)} \le 0\}}(\kappa^{(l)})$$

on $\mathbb{R}^k \times \mathbb{R}$ (with $\tau = \tau(\varepsilon) < 1$) such that for some $\rho^{(l)} \in \mathbb{R}$ and $K > 0$ large enough,

$$\mathbb{E}_Z \left[\left| \tilde{\delta}^{\varepsilon}(Z) - \rho^{(l)} - \int_{\|\kappa^{(l)}\| \le K} \sigma_g \left((\omega^{(l)})^\intercal Z + b^{(l)} \right) m(\kappa^{(l)}) \, d\kappa^{(l)} \right| \right] < 3\varepsilon.$$

Note that we can bound

$$\begin{aligned}
\left| \int_{\|\kappa^{(l)}\| \le K} m(\kappa^{(l)}) \, d\kappa^{(l)} \right| &= \left| \int_{\|\omega^{(l)}\| \le K} \frac{2\tau^{k+1}}{u_g - \ell_g} e^{-\pi^2\varepsilon^2\tau^2\|\omega^{(l)}\|^2} \int_{-\min\{R\|\omega^{(l)}\|, K - \|\omega^{(l)}\|\}}^{0} 2\pi \sin(2\pi b^{(l)}) \, db^{(l)} \, d\omega^{(l)} \right| \\
&= \left| \int_{\|\omega^{(l)}\| \le K} \frac{2\tau^{k+1}}{u_g - \ell_g} e^{-\pi^2\varepsilon^2\tau^2\|\omega^{(l)}\|^2} \left(\cos\left(2\pi \min\{R\|\omega^{(l)}\|, K - \|\omega^{(l)}\|\} \right) - 1 \right) \, d\omega^{(l)} \right| \\
&\le \tau \int_{\mathbb{R}^k} \frac{4}{u_g - \ell_g} e^{-\pi^2\varepsilon^2\tau^2\|\omega^{(l)}\|^2} \tau^k \, d\omega^{(l)} = \tau C_m < C_m.
\end{aligned}$$

For $\xi > 0$, consider the set

$$\mathcal{K}_\xi^{(l)} \triangleq \left\{ \kappa \in \mathbb{R}^{n \times k} \times \mathbb{R}^n \,\middle|\, \|\kappa^{(l)}\| \le K, \, \|\kappa^{(j)}\| \le \xi \text{ for } j \ne l \right\}.$$

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

$$m_\xi^{(l)}\left(\kappa^{(1)}, \cdots, \kappa^{(n)}\right) \triangleq m(\kappa^{(l)}) \frac{\mathbf{1}_{\mathcal{K}_\xi^{(l)}}\left(\kappa^{(1)}, \cdots, \kappa^{(n)}\right)}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)^{n-1}}.$$

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

$$\mathbb{E}_Z \left[\left\| \tilde{\delta}^{(\varepsilon,l)}(Z) - \mathbf{v}^{(\varepsilon,l)} - \int_{\mathbb{R}^{n\times k} \times \mathbb{R}^n} \phi(Z;\kappa) m_\xi^{(l)}(\kappa) \, d\kappa \right\|_2 \right] = \mathcal{O}(\varepsilon).$$

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

$$\tilde{\Phi}_\xi^{(l)}(z) \triangleq \int_{\mathbb{R}^{n\times k} \times \mathbb{R}^n} \phi(z;\kappa) m_\xi^{(l)}(\kappa) \, d\kappa = \begin{bmatrix} \int_{\mathbb{R}^{n\times k} \times \mathbb{R}^n} \sigma_g\left((\omega^{(1)})^\intercal z + b^{(l)} \right) m_\xi^{(l)}(\kappa) \, d\kappa \\ \vdots \\ \int_{\mathbb{R}^{n\times k} \times \mathbb{R}^n} \sigma_g\left((\omega^{(n)})^\intercal z + b^{(l)} \right) m_\xi^{(l)}(\kappa) \, d\kappa \end{bmatrix} \in \mathbb{R}^n.$$

For $j = 1, \dots, n$, we denote the j-th component function of $\tilde{\Phi}_\xi^{(l)}$ by $\left[\tilde{\Phi}_\xi^{(l)}\right]_j$. Observe that if we denote $d\kappa^{(-l)} = d\kappa^{(1)} \cdots d\kappa^{(l-1)} d\kappa^{(l+1)} \cdots d\kappa^{(n)}$, then by our construction of m and K,

$$\begin{aligned}\left[\tilde{\Phi}_\xi^{(l)}\right]_l(Z) &= \int_{\mathbb{R}^k \times \mathbb{R}} \sigma_g\left((\omega^{(l)})^\intercal Z + b^{(l)}\right) m(\kappa^{(l)}) \left(\int_{\mathbb{R}^{(n-1)\times k} \times \mathbb{R}^{n-1}} \frac{\mathbf{1}_{\mathcal{K}_\xi^{(l)}}\left(\kappa^{(1)}, \cdots, \kappa^{(n)}\right)}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)^{n-1}} d\kappa^{(-l)} \right) d\kappa^{(l)} \\ &= \int_{\|\kappa^{(l)}\| \le K} \sigma_g\left((\omega^{(l)})^\intercal Z + b^{(l)}\right) m(\kappa^{(l)})\, d\kappa^{(l)},\end{aligned}$$

which is 3ε-close to $\tilde{\delta}^\varepsilon(Z) - \rho^{(l)}$ within $L^1(q_Z(z)\,dz)$, regardless of ξ.

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

$$|\sigma_g(r) - \sigma_g(0)| < \frac{\varepsilon}{2(n-1)C_m} \tag{11}$$

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

$$\begin{aligned}\left[\tilde{\Phi}_\xi^{(l)}\right]_j(z) &= \int_{\mathbb{R}^{(n-1)\times k} \times \mathbb{R}^{n-1}} \sigma_g\left((\omega^{(j)})^\intercal z + b^{(j)}\right) \frac{\prod_{m \neq l} \mathbf{1}_{\mathcal{B}_\xi}(\kappa^{(m)})}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)^{n-1}} \left(\int_{\|\kappa^{(l)}\| \le K} m(\kappa^{(l)})\, d\kappa^{(l)} \right) d\kappa^{(-l)} \\ &= \frac{1}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)} \int_{\|\kappa^{(j)}\| \le \xi} \sigma_g\left((\omega^{(j)})^\intercal z + b^{(j)}\right) \left(\int_{\|\kappa^{(l)}\| \le K} m(\kappa^{(l)})\, d\kappa^{(l)} \right) d\kappa^{(j)}.\end{aligned}$$

Define

$$\rho^{(-l)} \triangleq \int_{\|\kappa^{(l)}\| \le K} \sigma_g(0) m(\kappa^{(l)})\, d\kappa^{(l)} = \frac{1}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)} \int_{\|\kappa^{(j)}\| \le \xi} \sigma_g(0) \left(\int_{\|\kappa^{(l)}\| \le K} m(\kappa^{(l)})\, d\kappa^{(l)} \right) d\kappa^{(j)}.$$

Then we have

$$\begin{aligned}\left| \left[\tilde{\Phi}_\xi^{(l)}\right]_j(z) - \rho^{(-l)} \right| &\le \frac{1}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)} \int_{\|\kappa^{(j)}\| \le \xi} \left| \sigma_g\left((\omega^{(j)})^\intercal z + b^{(j)}\right) - \sigma_g(0) \right| \left| \int_{\|\kappa^{(l)}\| \le K} m(\kappa^{(l)})\, d\kappa^{(l)} \right| d\kappa^{(j)} \\ &\le \frac{1}{\mathrm{Vol}\left(\mathcal{B}_\xi\right)} \int_{\|\kappa^{(j)}\| \le \xi} C_m \left| \sigma_g\left((\omega^{(j)})^\intercal z + b^{(j)}\right) - \sigma_g(0) \right| d\kappa^{(j)}.\end{aligned}$$

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

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

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

$$\left| \left[\tilde{\Phi}_\xi^{(l)}\right]_j(z) - \rho^{(-l)} \right| \le 2\|\sigma_g\|_\infty C_m$$

is enough, because $\mathrm{Prob}_Z[\|Z\| \ge R] < \frac{\varepsilon}{4(n-1)\|\sigma_g\|_\infty C_m}$. We have established

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

for all $j \neq l$.

Now, with

$$\mathbf{v}^{(\varepsilon,l)} = \begin{bmatrix} -\rho^{(-l)} \\ \vdots \\ -\rho^{(-l)} \\ \rho^{(l)} \\ -\rho^{(-l)} \\ \vdots \\ -\rho^{(-l)} \end{bmatrix} \leftarrow \text{on } l\text{-th coordinate},$$

we have

$$\begin{aligned}
&\mathbb{E}_Z\left[\left\|\tilde{\delta}^{(\varepsilon,l)}(Z) - \mathbf{v}^{(\varepsilon,l)} - \int_{\mathbb{R}^{n\times k}\times\mathbb{R}^n} \phi(Z;\kappa) m_\xi^{(l)}(\kappa)\, d\kappa\right\|_2\right] \\
&\le \mathbb{E}_Z\left[\left\|\tilde{\delta}^{(\varepsilon,l)}(Z) - \mathbf{v}^{(\varepsilon,l)} - \int_{\mathbb{R}^{n\times k}\times\mathbb{R}^n} \phi(Z;\kappa) m_\xi^{(l)}(\kappa)\, d\kappa\right\|_1\right] \\
&= \mathbb{E}_Z\left[\left|\tilde{\delta}^{\varepsilon}(Z) - \rho^{(l)} - \left[\tilde{\Phi}_\xi^{(l)}\right]_l(z)\right|\right] + \sum_{j\neq l}\mathbb{E}_Z\left[\left|\left[\tilde{\Phi}_\xi^{(l)}\right]_j(Z) - \rho^{(-l)}\right|\right] \\
&< 3\varepsilon + (n-1)\frac{\varepsilon}{n-1} = 4\varepsilon.
\end{aligned}$$

The space of vector functions $h = ([h]_1, \ldots, [h]_n)\colon \mathbb{R}^k \to \mathbb{R}^n$ 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 $\langle g, h\rangle_{\mathcal{H}} = \sum_{j=1}^n \mathbb{E}_Z[[g]_j(Z)[h]_j(Z)] = \mathbb{E}_Z[g^\intercal(Z)h(Z)]$. Now let $p(\kappa) > 0$ be the density function on $\mathbb{R}^{n\times k}\times\mathbb{R}^n$ from which we sample κ's, and define

$$C_{\mathcal{K}}^{(l)} \triangleq \sup_\kappa \frac{m_\xi^{(l)}(\kappa)}{p(\kappa)},$$

which is finite because $m_\xi^{(l)}$ is bounded and compactly supported, while p is positive and continuous. For each random κ_i, $i = n+1, \ldots, N_g$, the corresponding realization

$$h_i := \frac{m_\xi^{(l)}(\kappa_i)}{p(\kappa_i)}\,\phi(\cdot;\kappa_i) \in \mathcal{H}$$

satisfies $\|h_i\|_{\mathcal{H}} \le \sqrt{n}\|\sigma_g\|_\infty C_{\mathcal{K}}^{(l)}$. Hence, as in the $n = 1$ case,

$$\left\|\frac{1}{N_g - n}\sum_{i=n+1}^{N_g} h_i - \int_{\mathbb{R}^{n\times k}\times\mathbb{R}^n}\phi(Z;\kappa)m_\xi^{(l)}(\kappa)\, d\kappa\right\|_{\mathcal{H}} \le \frac{\sqrt{n}\|\sigma_g\|_\infty C_{\mathcal{K}}^{(l)}}{\sqrt{N_g - n}}\left(1 + \sqrt{2\log\frac{1}{\zeta}}\right)$$

with probability $\ge 1 - \zeta$ over $\kappa_{n+1}, \ldots, \kappa_{N_g}$. Let $\theta_i^{(\varepsilon,l)} = \frac{m_\xi^{(l)}(\kappa_i)}{(N_g - n)p(\kappa_i)}$ 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 $b_1, \ldots, b_n \in \mathbb{R}^n$, in a way that the constant vectors $\sigma_g(b_i)$ are linearly independent. Then there exist $\theta_1^{(\varepsilon,l)}, \ldots, \theta_n^{(\varepsilon,l)} \in \mathbb{R}$ such that

$$\sum_{i=1}^n \theta_i^{(\varepsilon,l)}\sigma_g(b_i) = \sum_{i=1}^n \theta_i^{(\varepsilon,l)}\phi(z;\kappa_i) = \mathbf{v}^{(\varepsilon,l)}.$$

Given $f\colon \mathbb{R}^k \to \mathbb{R}^n$, let $M = \sup_{z\in\mathbb{R}^k} \|f(z)\|_2$. Chaining all the approximation steps, we have

$$\begin{aligned}
&\left|\mathbb{E}_Z\left[\left(\delta^{(l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(\varepsilon,l)}\phi(Z;\kappa_i)\right)^\intercal f(Z)\right]\right| \\
&\le \left|\mathbb{E}_Z\left[\left(\delta^{(l)}(Z) - \tilde{\delta}^{(\varepsilon,l)}(Z)\right)^\intercal f(Z)\right]\right| + \mathbb{E}_Z\left[\left|\left(\tilde{\delta}^{(\varepsilon,l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(\varepsilon,l)}\phi(Z;\kappa_i)\right)^\intercal f(Z)\right|\right] \\
&\le \left|\mathbb{E}_Z\left[\left(\delta(Z) - \tilde{\delta}^{\varepsilon}(Z)\right)[f]_l(Z)\right]\right| + M\,\mathbb{E}_Z\left[\left\|\delta^{(\varepsilon,l)}(Z) - \sum_{i=1}^{N_g} \theta_i^{(\varepsilon,l)}\phi(Z;\kappa_i)\right\|_2\right] \\
&\le C_\delta\,\varepsilon \sup_{z\in\mathbb{R}^k}\left(\left|[f]_l(z)\right| + \left\|\nabla[f]_l\right\|\right) + M\,\mathbb{E}_Z\left[\left\|\tilde{\delta}^{(\varepsilon,l)}(Z) - \mathbf{v}^{(\varepsilon,l)} - \int_{\mathbb{R}^{n\times k}\times\mathbb{R}^n} \phi(Z;\kappa) m_\xi^{(l)}(\kappa)\,d\kappa\right\|_2\right] \\
&\quad + M\,\mathbb{E}_Z\left[\left\|\mathbf{v}^{(\varepsilon,l)} + \int_{\mathbb{R}^{n\times k}\times\mathbb{R}^n} \phi(Z;\kappa) m_\xi^{(l)}(\kappa)\,d\kappa - \sum_{i=1}^{N_g} \theta_i^{(\varepsilon,l)}\phi(Z;\kappa_i)\right\|_2\right] \\
&\le \left((C_\delta+4)\varepsilon + \frac{\sqrt{n}\|\sigma_g\|_\infty C_{\mathcal{K}}^{(l)}}{\sqrt{N_g - n}}\left(1 + \sqrt{2\log\frac{1}{\zeta}}\right)\right) \sup_{z\in\mathbb{R}^k}\left(\|f(z)\|_2 + \|Df(z)\|\right)
\end{aligned}$$

with probability $\ge 1-\zeta$. 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

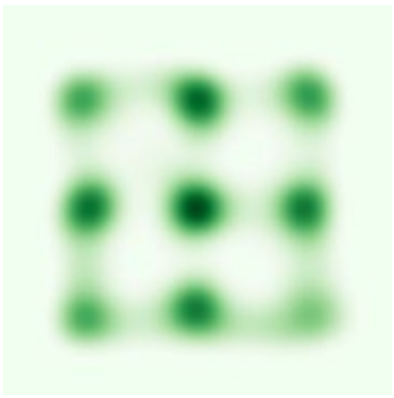
We now provide details of the experiments for Figure 6. The code is available at `https://github.com/sehyunkwon/Infinite-WGAN`. The true and latent distributions are 2-dimensional, i.e., $n = 2$ and $k = 2$. The true distribution P_X is a mixture of 8 Gaussians with equal weights, where the means are $(\sqrt{2}\cos\frac{m\pi}{4}, \sqrt{2}\sin\frac{m\pi}{4})$ 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 $\phi_i(x) = \sigma_g(\kappa_w z + \kappa_b)$ as described in (AG) for $1 \le i \le N_g = 5000$, where the activation function σ_g is $\tanh$. Weights κ_w and κ_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 (κ_w, κ_b) with $(\mathbf{0}_{2\times 2}, (1,0))$ and $(\mathbf{0}_{2\times 2}, (0,1))$. The discriminator feature functions are of form $\psi_j(x) = \sigma(a_j^\intercal x + b_j)$ as described in (AD) for $1 \le j \le N_d = 1000$ where the activation function σ 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, $\frac{x}{\tilde{a}} + \frac{y}{\tilde{b}} = 1$ 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 $b_j = -c$.

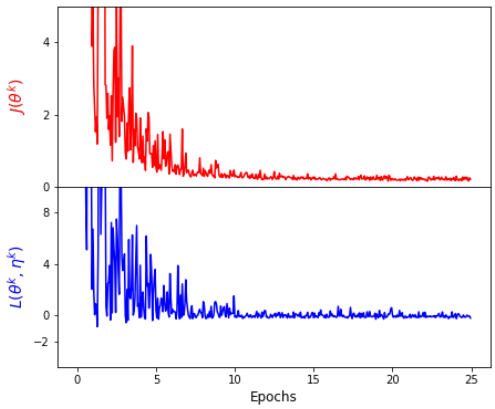
The generator stepsize starts at $\alpha = 10^{-5}$ and decays by a factor of 0.9 at every epoch. The networks are trained for 25 epochs with (X-sample) batch size 5000. At each iteration, 5000 latent vectors (Z-samples) are sampled IID from the standard Gaussian distribution for the stochastic gradient ascent step, and another 5000 latent vectors are sampled IID from the standard Gaussian distribution for the stochastic gradient descent step. The generator parameter θ is randomly initialized (IID) with the Gaussian distribution with zero mean and variance 5×10^{-3}. The discriminator parameter γ 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

(a) Samples from true distribution P_X



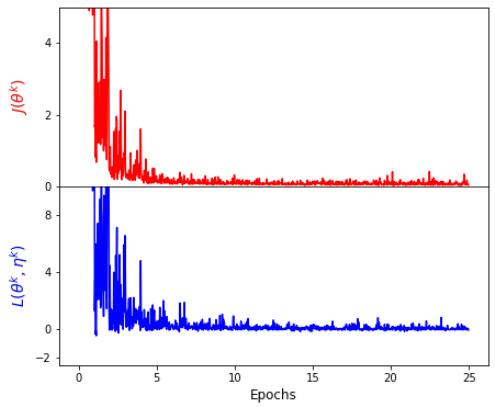
(b) Samples from generator $g_\theta(Z)$



(c) Convergence of the loss functions J and L

(d) Samples from true distribution P_X

(e) Samples from generator $g_\theta(Z)$



(f) Convergence of the loss functions J and L

Figure 6. Additional experiments with mixtures of Gaussians. The code is available at https://github.com/sehyunkwon/Infinite-WGAN.

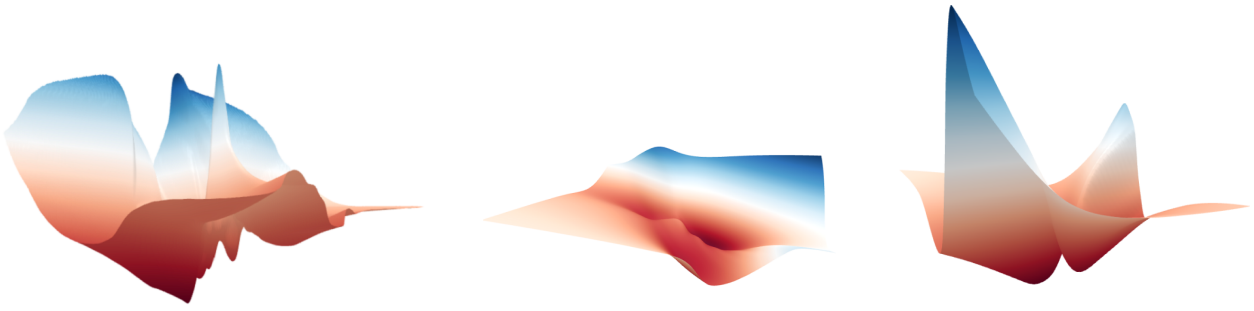
$N_g = 10,000$ for generator feature functions. The generator parameter θ is randomly initialized (IID) with the Gaussian distribution with zero mean and variance 3×10^{-3}. Discriminator feature functions are generated in the same manner. Figure 6(a) shows the true distribution, and Figure 6(b) shows 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 $(\frac{m}{20}\cos\frac{2m}{20}\pi, \frac{m}{20}\sin\frac{2m}{20}\pi)$ 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 θ is randomly initialized (IID) with the Gaussian distribution with zero mean and variance 3×10^{-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 6(d) shows the true distribution, and Figure 6(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

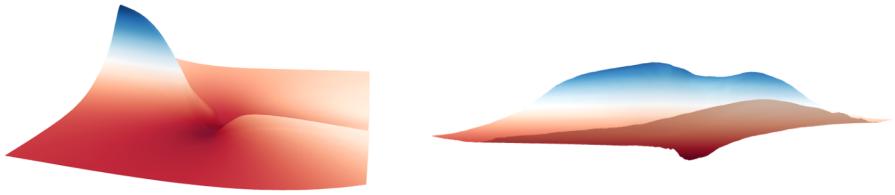
In this section, we describe the experiments for Figure 5, which visualizes the loss landscape of $J(\theta)$ for the cases $N_g = 2$ and $N_g = 10$. We also provide additional experiments for $N_g = 3$, 5, and 100. In the $N_g = 2$ case, the landscape is highly non-convex and displays at least three non-global local minima. We observe that in Figures 7 and 8, 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 σ_g is $\tanh$. Weights κ_w are randomly sampled (IID) from an isotropic Gaussian and then multiplied by a scalar factor, sampled independently from the standard normal distribution. Weights κ_b are randomly sampled (IID) from the Gaussian distribution with zero mean and variance 3×10^{-1}. The discriminator feature functions are of the form $\psi_j(x) = \sigma(a_j^\intercal x + b_j)$ as described in (AD) for $1 \leq j \leq N_d = 8$, where the activation function σ is $\tanh$.

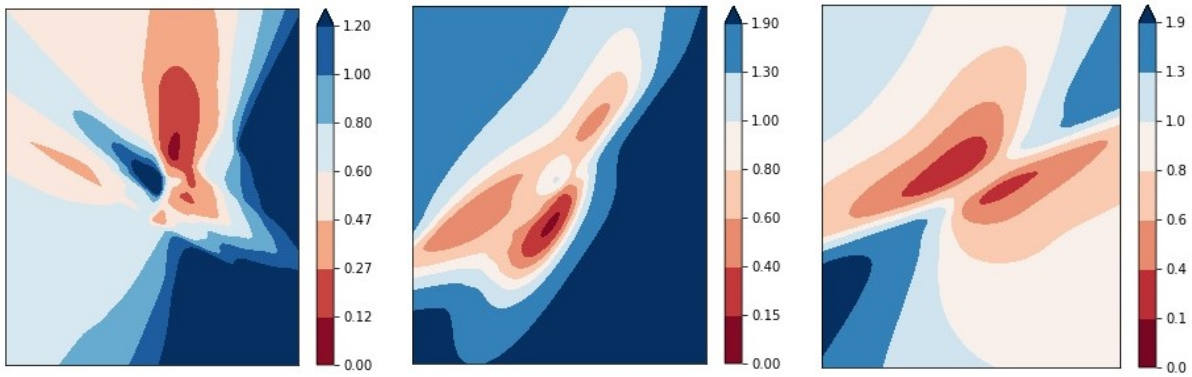


(a) $N_g = 2$ **(b)** $N_g = 3$ **(c)** $N_g = 5$

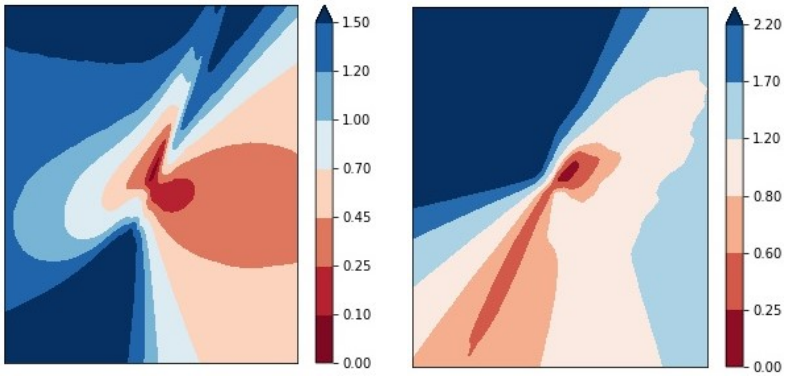


(d) $N_g = 10$ **(e)** $N_g = 100$

Figure 7. Loss landscapes of $J(\theta)$ for $N_g = 2, 3, 5, 10$, and 100.



(a) $N_g = 2$ **(b)** $N_g = 3$ **(c)** $N_g = 5$



(d) $N_g = 10$ **(e)** $N_g = 100$

Figure 8. Corresponding contour plots of the landscapes of Figure 7.