A Useful Mathematical Results

Theorem A.1. *Let A be $m \times m$ random matrix whose entries A_{ij} are independent identically distributed standard Gaussian random variables. Then, there exists absolute constant $c, C > 0$ such that*

$$\|A\|_{op} \leq C\sqrt{m}, \quad \textit{with probability at least } 1 - 2e^{-cm}. \tag{16}$$

Theorem A.2 (Strong Bai-Yin theorem)**.** *Let A be $m \times m$ random matrix whose entries A_{ij} are independent identically distributed standard Gaussian random variables. Then*

$$\lim_{m\to\infty} \|A\|_{op}/\sqrt{m} = \sqrt{2}, \quad \textit{almost surely.} \tag{17}$$

Theorem A.3 (Kolmogorov's SLLN for *i.i.d.*)**.** *Let $\{X_n\}$ be sequence of i.i.d. random variables and $S_n = \sum_{i=1}^{n} X_i$. Then $\frac{S_n}{n} \overset{a.s.}{\to} \mathbb{E}X_1$ if and only if $\mathbb{E}|X_1| < \infty$.*

Lemma A.1 (Almost surely convergence)**.** *Some important properties of almost surely convergence.*

1. *If $X_n \overset{a.s.}{\to} X$, then $g(X_n) \overset{a.s.}{\to} g(X)$ for all continuous function g.*
2. *If $X_n \overset{a.s.}{\to} X$ and $Y_n \overset{a.s.}{\to} Y$, then $X_nY_n \overset{a.s.}{\to} XY$.*
3. *If $X_n \overset{a.s.}{\to} X$ and $Y_n \overset{a.s.}{\to} Y$, then $aX_n + bY_n \overset{a.s.}{\to} aX + bY$.*

Lemma A.2 (Gaussian smoothing)**.** *Let f, g be a real-valued function. Define function $F(\sigma) := \mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)} f(z)$ and $G(\mu) = \mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)} g(z)$ for $\sigma > 0$. Suppose $f(x), g(x) \in o(e^{-x^2})$, then*

$$F'(\sigma) = \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[f(\mu+\sigma z)(z^2-1)\right]$$
$$G'(\mu) = \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g(\mu+\sigma z)z\right]$$

Proof. Note that $F(\sigma) = \mathbb{E}_{z\sim\mathcal{N}(0,1)} f(\mu + \sigma z)$, then

$$\begin{aligned}
F'(\sigma) =& \frac{d}{d\sigma}\int_{-\infty}^{\infty} f(\mu+\sigma z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
=& \int_{-\infty}^{\infty} f'(\mu+\sigma z) z \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
=& \int_{-\infty}^{\infty} f'(u)\left(\frac{u-\mu}{\sigma}\right)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(u-\mu)^2}{2\sigma^2}}du, \quad u = \mu+\sigma z \\
=& \frac{1}{\sigma}\int_{-\infty}^{\infty} f(u)\left[\frac{(u-\mu)^2}{\sigma^2}-1\right]\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(u-\mu)^2}{2\sigma^2}}du \\
=& \frac{1}{\sigma}\int_{-\infty}^{\infty} f(\mu+\sigma z)\left[z^2-1\right]\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \\
=& \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[f(\mu+\sigma z)(z^2-1)\right]
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
G'(\mu) =& \frac{d}{d\mu}\int_{-\infty}^{\infty} g(\mu+\sigma z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
=& \int_{-\infty}^{\infty} g'(\mu+\sigma z)\frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \\
=& \int_{-\infty}^{\infty} g'(u)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(u-\mu)^2}{2\sigma^2}}du, \quad u=\mu+\sigma z \\
=& \frac{1}{\sigma}\int_{-\infty}^{\infty} g(u)\left(\frac{u-\mu}{\sigma}\right)\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(u-\mu)^2}{2\sigma^2}}du \\
=& \frac{1}{\sigma}\int_{-\infty}^{\infty} g(\mu+\sigma z)z\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \\
=& \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g(\mu+\sigma z)z\right]
\end{aligned}$$

□

Lemma A.3 (Gaussian conditioning)**.** *Given $G\in\mathbb{R}^{n\times m}$ and $H\in\mathbb{R}^{n\times m}$, let $W\in\mathbb{R}^{n\times n}$ to follow matrix Gaussian distribution, i.e., $W\sim\mathcal{MN}(0,\sigma I_n,\sigma I_n)$ for some $\sigma>0$, suppose $G=WH$ has feasible solutions. Then the conditional distribution of W given on $G=WH$ is*

$$W|_{G=WH}\sim\mathcal{MN}(GH^\dagger, I_n, \sigma^2\Pi\Pi^T).$$

where $\Pi=I_n-HH^\dagger$ is the orthogonal projection onto the null(H^T).

Proof. First, we consider the optimization problem

$$\min_W \frac{1}{2}\|W\|_F^2, \quad s.t. \quad G=WH.$$

The Lagrange function is given by

$$L(W,V)=\frac{1}{2}\|W\|_F^2+\langle V, G-WH\rangle.$$

The KKT condition implies $\nabla_W L(W,V)=W-VH^T=0$ and further $W=VH^T$. Since $G=WH$, we have $V=G(H^TH)^\dagger$ and so $W^*=G(H^TH)^\dagger H^T=GH^\dagger$.

Then let $\Pi=I_n-HH^\dagger$ be the orthogonal projection onto the null(H^T). Thus, the conditional distribution of W given $G=WH$ is

$$W|_{G=WH}=GH^\dagger+\tilde{W}\Pi^T=\mathcal{MN}(GH^\dagger, I_n, \sigma^2\Pi\Pi^T).$$

□

Lemma A.4 (Conditional distribution)**.** *Let $X\sim\mathcal{MN}(M,U,V)$. Partition X, M, and V such that*

$$X=\begin{bmatrix}X_1\\X_2\end{bmatrix}, \quad M=\begin{bmatrix}M_1\\M_2\end{bmatrix} \quad U=\begin{bmatrix}U_{11} & U_{12}\\U_{21} & U_{22}\end{bmatrix}$$

where $X_1\in\mathbb{R}^{m\times p}$. Then

$$\begin{aligned}
&X_1\sim\mathcal{MN}(M_1,U_{11},V)\\
&X_2|X_1\sim\mathcal{MN}\left(M_2+U_{21}U_{11}^{-1}(X_1-M_2), U_{22}-U_{21}U_{11}^{-1}U_{12}, V\right).
\end{aligned}$$

*Note, if $U_{21}=0$, then $X_2|X_1\sim\mathcal{MN}(M_2,U_{22},V)$ indicates X_2 and X_1 are **independent**.*

Lemma A.5. *Let σ be a L-Lipschitz continuous function. Then σ is also a controllable function. In addition, $\phi(x,y):=\sigma(x)\sigma(y)$ is also a controllable function.*

Proof. WOLG, we can assume $L = 1$. As σ is Lipschitz continuous on its region, there must exists some x_0 such that $\sigma(x_0) = c$. Then we have

$$|\sigma(x)| \leq |\sigma(x) - \sigma(x_0)| + |c| \leq |x - x_0| + |c| \leq e^{|c|^{-1}|x-x_0|} \leq e^{|c|^{-1}|x_0|} e^{|c|^{-1}|x|} = C_1 e^{C_2|x|}.$$

Similarly, we have

$$|\phi(x, y)| = |\sigma(x)|\, |\sigma(y)| \leq C_1 e^{C_2(|x|+|y|)}.$$

$\square$

Lemma A.6. *Let f be a controllable function. Then for all $\mu \in \mathbb{R}$ and $\sigma \geq 0$, we have*

$$\mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)}\, |f(z)| \leq 2C_1 e^{C_2|\mu| + C_2^2\sigma^2/2}.$$

Proof. Note that

$$\begin{aligned}
\mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)}\, |f(z)| =& \mathbb{E}_{z\sim\mathcal{N}(0,1)}\, |f(\sigma z + \mu)| \\
\leq& \mathbb{E}_{z\sim\mathcal{N}(0,1)} C_1 e^{C_2(\sigma|z| + |\mu|)} \\
=& C_1 e^{C_2|\mu|} \mathbb{E}_{z\sim\mathcal{N}(0,1)} e^{C_2\sigma|z|} \\
=& C_1 e^{|\mu|} \int_{-\infty}^{\infty} e^{C_2\sigma|z|} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \\
=& C_1 e^{|\mu|} \left[\int_{-\infty}^{0} e^{-C_2\sigma z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz + \int_{0}^{\infty} e^{C_2\sigma z} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \right] \\
=& C_1 e^{|\mu|} \left[\int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z + C_2\sigma)^2 + \frac{C_2^2\sigma^2}{2}} dz + \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z - C_2\sigma)^2 + \frac{C_2^2\sigma^2}{2}} dz \right] \\
\leq& 2C_1 e^{C_2|\mu| + C_2^2\sigma^2/2}
\end{aligned}$$

$\square$

B Proof of Theorem 4.1

In this Appendix, we show the preactivation g_k^ℓ acts like Gaussian random variable. As a consequence, the finite-depth neural network f_θ^L tends to a Gaussian process as width $n \to \infty$.

Lemma B.1. *Suppose the activation function ϕ is nonlinear Lipschitz continuous function. For input x, let $g^1, \cdots, g^\ell$ be the resulting pre-activations for $\ell \in [L]$. Then for any $\ell \in [L]$ and for any controllable function $\Phi : \mathbb{R}^\ell \to \mathbb{R}$, we have as $m \to \infty$*

$$\frac{1}{n}\sum_{k=1}^{n} \Phi(g_k^1, \cdots, g_k^\ell) \xrightarrow{a.s.} \mathbb{E}\left[\Phi(z^1, \cdots, z^\ell)\right], \tag{18}$$

where $(z^i, z^j) \sim \mathcal{N}(0, \Sigma)$ and the covariance matrix $\Sigma \in \mathbb{R}^{2\times 2}$ are computed recursively as follows

$$\Sigma(z^1, z^i) = \delta_{1,i}\sigma_u^2\|x\|^2/n_{in}, \qquad \forall i \geq 1, \tag{19}$$

$$\Sigma(z^i, z^j) = \sigma_w^2 \mathbb{E}\phi(u^{i-1})\phi(u^{j-1}), \qquad \forall i \geq 2. \tag{20}$$

where $u^1 = z^1$ and $u^\ell = z^\ell + z^1$ with covariance

$$\Sigma(u^1, u^i) = \sigma_u^2\|x\|^2/n_{in}, \qquad \forall i \geq 1, \tag{21}$$

$$\Sigma(u^i, u^j) = \Sigma(z^i, z^j) + \Sigma(z^1, z^1), \qquad \forall i \geq 2. \tag{22}$$

If, in addition, W^i and W^j are independent, then

$$\Sigma(z^i, z^j) = 0, \quad \forall i \neq j. \tag{23}$$

Lemma B.2. *[37, Theorem 5.4] For any NETSOR program whose weight matrices are random initiated as* (5) *and all activation functions are controllable. If $g^1, \cdots, g^\ell$ are any G-vars (i.e., pre-activation in our case), then for any controllable function $\Phi : \mathbb{R}^\ell \to \mathbb{R}$, we have*

$$\frac{1}{n}\sum_{k=1}^{n} \Phi(g_k^1, \cdots, g_k^\ell) \xrightarrow{a.s.} \mathbb{E}_{z\sim\mathcal{N}(\mu,\Sigma)}\Phi(z), \tag{24}$$

where $z := (z^1, \cdots, z^\ell)$ and μ and Σ can computed by [37, Definition 5.2].

Intuitively, these two Lemmas indicate that $(g_k^1, \cdots, g_k^\ell)$ acts like a multidimensional Gaussian vector whose covariance can be computed recursively. Lemma B.1 is a special case of Lemma B.2 as Lemma B.1 requires each pre-activation g^ℓ encoded same input x, while Lemma B.2 does not make such assumption. In fact, the proof techniques are identical, *i.e.*, uses Gaussian conditions and smoothing inductively on previous results. To make the paper self-contained, here we provide a proof for Lemma B.1 where we simplify the proof of [37, Theorem 5.4] in the following subsections by removing so-called *core set*.

B.1 Proof of Theorem 4.1 by Using Master Theorem B.1 or B.2

Based on Lemma B.1 or B.2, we can immediately obtain the desired result.

For simplicity, we assume $\sigma_\ell = 1$. We prove the desired result by induction. For $L = 1$, we have $f_\theta^L(x) = g^1(x) = W^1 x$ and

$$f_{\theta,k}^1(x) = g_k^1(x) = \langle w_k, x\rangle \overset{i.i.d.}{\sim} \mathcal{N}(0, \|x\|^2/n_{in}).$$

Then we have

$$\hat{\Sigma}^1(x, x') = \mathrm{cov}(f_{\theta,k}^L(x), f_{\theta,k}^L(x')) = \langle x, x'\rangle := \Sigma^1(x, x').$$

For $L = 2$, we have $f_\theta^L(x) = g^2(x) = W^2 h^1(x)$. By condition on g^1, we have

$$f_{\theta,k}^2(x) = g_k^2(x) = \left\langle w_k^2, h^1(x)\right\rangle \overset{i.i.d.}{\sim} \mathcal{N}(0, \|h^1(x)\|^2/n).$$

Then

$$\begin{aligned}\hat{\Sigma}^2(x,x') &= \left\langle h^1(x), h^1(x')\right\rangle /n \\ &= \left\langle \phi(g^1(x)), \phi(g^1(x'))\right\rangle /n \\ &= \frac{1}{n}\sum_{i=1}^{n}\phi(g_i^1(x))\phi(g_i^1(x')) \\ &\xrightarrow{a.s.} \mathbb{E}\phi(z^1(x))\phi(z^2(x')) \\ &=: \Sigma^2(x,x'),\end{aligned}$$

where

$$(z^1(x), z^1(x')) \sim \mathcal{N}\left(0, \begin{bmatrix}\Sigma^1(x,x) & \Sigma^1(x,x') \\ \Sigma^1(x',x) & \Sigma^1(x',x')\end{bmatrix}\right).$$

Now, we assume the results holds for L. Then we show the result for $f_\theta^{L+1}(x)$. In this case, we have $f_\theta^{L+1}(x) = g^{L+1}(x)$. By condition on the values g^L, we have the output $f_{\theta,k}^{L+1}$ are *i.i.d.*centered Gaussian random variables, *i.e.*,

$$f_{\theta,k}^{L+1}(x) = g_k^{L+1}(x) = \left\langle w_k^{L+1}, h^L(x)\right\rangle \overset{i.i.d.}{\sim} \mathcal{N}(0, \|h^L(x)\|^2/n).$$

Then we have

$$\begin{aligned}\hat{\Sigma}^{L+1}(x,x') &= \mathrm{Cov}(f_{\theta,k}^{L+1}(x), f_{\theta,k}^{L+1}(x')) \\ &= \left\langle h^L(x), h^L(x')\right\rangle /n \\ &= \frac{1}{n}\sum_{i=1}^{n}\phi(g_i^L(x)+g_i^1(x))\phi(g_i^L(x')+g_i^1(x')) \\ &\xrightarrow{a.s.} \mathbb{E}\phi(z^L(x)+z^1(x))\phi(z^L(x')+z^1(x')) \\ &=: \Sigma^{L+1}(x,x').\end{aligned}$$

where

$$\begin{bmatrix} z^1(x) \\ z^L(x) \\ z^1(x') \\ z^L(x') \end{bmatrix} \sim \mathcal{N}\left(0, \left[\begin{array}{cc|cc} \Sigma^1(x,x) & 0 & \Sigma^1(x,x') & 0 \\ 0 & \Sigma^L(x,x) & 0 & \Sigma^L(x,x') \\ \hline \Sigma^1(x',x) & 0 & \Sigma^1(x',x') & 0 \\ 0 & \Sigma^L(x',x) & 0 & \Sigma^L(x',x') \end{array}\right]\right).$$

Here the covariance is deterministic and independent of g^L. Consequently, the conditioned and unconditioned distributions of $f_{\theta,k}^{L+1}$ are equal in the limit: they are *i.i.d.*centered Gaussian random variables with covariance Σ^{L+1}.

B.2 Proof of Lemma B.1: the basic case $\ell = 1$

WLOG, we can assume $\sigma_\ell = 1$. We prove by induction. When $\ell = 1$, we have

$$g^1 = W_1 x$$

so that

$$g_k^1 \overset{i.i.d.}{\sim} \mathcal{N}(0, \|x\|^2/n_{in}).$$

Given a controllable function Φ, the random variables $X_k = \Phi(g_k^1)$ are still *i.i.d.*. It follows from Lemma A.6 that

$$\mathbb{E}\left|X_1\right| = \mathbb{E}_{z\sim\mathcal{N}(0,\|x\|^2)}\left|\Phi(z)\right| \le C_1 e^{C_2^2\|x\|^2} < \infty.$$

Then the desired result is obtained by following Theorem A.3 the classical Kolmogorov's SLLN for $i.i.d.$ random variables.

B.3 Proof of Lemma B.1: general case for independent matrices $W^k \neq W^\ell$

Suppose the desired result hold for ℓ, then we show the result also hold for $\ell+1$. In addition, we assume the weight matrices W^ℓ are independent to each other. Thus, the weight matrix $W^{\ell+1}$ are not used in previous layers. For brevity, we denote $W := W^{\ell+1}$ and so we have expression

$$g^{\ell+1} = Wh^\ell.$$

Here the randomness of $g^{\ell+1}$ comes from both W and h^ℓ. As W is not used before, W and h^ℓ are independent. Let $\mathcal{B}$ be the σ-algebra spanned by all previous $g^1, g^2, \cdots, g^\ell$. Then the conditional distribution of $g^{\ell+1}$ on $\mathcal{B}$ is given by

$$g^{\ell+1}|\mathcal{B} \sim \mathcal{N}(0, \|h^\ell\|^2/nI_n),$$

or equivalently

$$g_k^{\ell+1}|\mathcal{B} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \|h^\ell\|^2/n). \tag{25}$$

By using the inductive hypothesis, we have

$$\sigma_n^2 := \|h^\ell\|^2/n = \frac{1}{n}\sum_{k=1}^n \phi(g_k^\ell + g_k^1)^2 \overset{a.s.}{\to} \mathbb{E}\left[\phi(z^\ell + z^1)\right]^2 = \Sigma(z^{\ell+1}, z^{\ell+1}) := \sigma^2, \tag{26}$$

where we use the fact $\Phi(x,y) := \phi(x+y)$ is controllable, *i.e.*,

$$|\Phi(x,y)| = |\phi(x+y)| \leq |x+y| \leq e^{|x|+|y|}.$$

By using triangle inequality, we have

$$\left|\frac{1}{n}\sum_{k=1}^n \Phi(g_k^1, \cdots, g_k^{\ell+1}) - \mathbb{E}\left[\Phi(z^1, \cdots, z^{\ell+1})\right]\right| \leq |A_n| + |B_n| + |C_n|,$$

where

$$A_n = \frac{1}{n}\sum_{k=1}^n \Phi(g_k^1, \cdots, g_k^{\ell+1}) - \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) \tag{27}$$

$$B_n = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) - \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) \tag{28}$$

$$C_n = \frac{1}{n}\sum_{k=1}^n \Phi(g_k^1, \cdots, g_k^\ell, z) - \mathbb{E}\left[\Phi(z^1, \cdots, z^{\ell+1})\right] \tag{29}$$

A_n converges to 0 almost surely

Define random variables $Z_k := \Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z)$. By equation (25), we have $g_k^{\ell+1}|\mathcal{B} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_n^2)$, we can easily show X_k are centered and uncorrelated. Observe that

$$\begin{aligned}
\mathbb{E}Z_k =& \mathbb{E}_\mathcal{B}\mathbb{E}_{g^{\ell+1}|\mathcal{B}} Z_k \\
=& \mathbb{E}_\mathcal{B}\mathbb{E}_{g^{\ell+1}|\mathcal{B}}\left[\Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z)\right] \\
=& \mathbb{E}_\mathcal{B}\left[\mathbb{E}_{g^{\ell+1}|\mathcal{B}}\Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z)\right] \\
=& \mathbb{E}_\mathcal{B}\left[\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z)\right] \\
=& \mathbb{E}_\mathcal{B}\left[0\right] = 0.
\end{aligned}$$

Similarly, we obtain $\mathbb{E}Z_kZ_{k'} = \delta_{kk'}\mathbb{E}\left|Z_k\right|^2$. Moreover, we can upper bound $\mathbb{E}\left[Z_k|\mathcal{B}\right]^2$ as follows

$$
\begin{aligned}
\mathbb{E}\left[Z_k|\mathcal{B}\right]^2 =& \mathbb{E}_{g^{\ell+1}|\mathcal{B}}\left|\Phi(g_k^1,\cdots,g_k^\ell,g_k^{\ell+1}) - \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\Phi(g_k^1,\cdots,g_k^\ell,z)\right|^2 \\
\le& 8\mathbb{E}_{z\sim\mathcal{N}(0,\sigma_n^2)}\left|\Phi(g_k^1,\cdots,g_k^\ell,z)\right|^2, \quad (a) \\
=& 8\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left|\Phi(g_k^1,\cdots,g_k^\ell,\sigma_n z)\right|^2 \\
\le& 8\mathbb{E}_{z\sim\mathcal{N}(0,1)}C_1 e^{2C_2(\sum_{i=1}^\ell |g_k^i| + \sigma_n|z|)}, \quad \Phi \text{ is controllable} \\
=& 8C_1 e^{2C_2\sum_{i=1}^\ell |g_k^i|}\mathbb{E}_{z\sim\mathcal{N}(0,1)}e^{2C_2\sigma_n|z|} \\
\le& 8C_1 e^{2C_2\sum_{i=1}^\ell |g_k^i|}e^{2C_2^2\sigma_n^2}.
\end{aligned}
$$

where (a) is due to maximal and Jensen's inequality.

Since $e^{2C_2\sum_{i=1}^\ell |g_k^i|}$ is controllable and $\sigma_n \xrightarrow{a.s.} \sigma$, it follows from the inductive hypothesis that

$$
\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[Z_k|\mathcal{B}\right]^2 \le 8C_1\cdot\left(\frac{1}{n}\sum_{k=1}^n e^{2C_2\sum_{i=1}^\ell |g_k^i|}\right)\cdot e^{2C_2^2\sigma_n^2} \xrightarrow{a.s.} 8C_1\mathbb{E}e^{2C_2\sum_{i=1}^\ell |z_i|}\cdot e^{2C_2^2\sigma_2^2}.
$$

As the RHS is a deterministic constant, we have

$$
\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[Z_k|\mathcal{B}\right]^2 \in o(n^\rho), \quad \forall\rho > 0.
$$

or equivalently, $\frac{1}{n}\sum_{k=1}^n \mathbb{E}[Z_k|\mathcal{B}]^2 \le n^\rho$ for large enough n.

Now, we will first show $A_{n^2} \xrightarrow{a.s.} 0$. For any $\epsilon > 0$, we have for large enough n

$$
\begin{aligned}
\mathbb{P}(|A_{n^2}| \ge \epsilon) \le& \epsilon^{-2}n^{-4}\mathbb{E}\left|A_{n^2}\right|^2 \\
=& \epsilon^{-2}n^{-4}\sum_{k,k'=1}^{n^2}\mathbb{E}[Z_kZ_{k'}] \\
=& \epsilon^{-2}n^{-4}\sum_{k=1}^{n^2}\mathbb{E}\left|Z_k\right|^2 \\
=& \epsilon^{-2}n^{-2}\mathbb{E}_{\mathcal{B}}\left[\frac{1}{n^2}\sum_{k=1}^{n^2}\mathbb{E}\left|Z_k|\mathcal{B}\right|^2\right] \\
=& \epsilon^{-2}n^{-2}\mathbb{E}_{\mathcal{B}}\left[n^{2\rho}\right] \\
\le& \epsilon^{-2}n^{-2+2\rho}.
\end{aligned}
$$

Furthermore, we obtain

$$
\sum_{n=1}^\infty \mathbb{P}(|A_{n^2}| \ge \epsilon) \le \sum_{n=1}^\infty \epsilon^{-2}n^{-2+2\rho} < \infty,
$$

provided we choose $0 < \rho < 1/2$. Thus, it follows from Borel-Cantelli lemma that $A_{n^2} \xrightarrow{a.s.} 0$.

Now for each n, we define $k_n := \sup\{k \in \mathbb{N} : k^2 \le n\}$, then we have $k_n^2 \le n \le (k_n+1)^2$. Note that

$$
A_n = \frac{1}{n}\sum_{i=1}^n Z_i = \frac{1}{n}\sum_{i=1}^{k_n^2} Z_i + \frac{1}{n}\sum_{i=k_n^2+1}^n Z_i.
$$

We will show the two terms goes 0 a.s.. As we just proved, the first term goes to 0 a.s., since

$$
\left|\frac{1}{n}\sum_{i=1}^{k_n^2} Z_i\right| \le \left|\frac{1}{k_n^2}\sum_{i=1}^{k_n^2} Z_i\right| \xrightarrow{a.s.} 0.
$$

For the second term, let $T_n := \frac{1}{n}\sum_{i=k_n^2+1}^{n} Z_i$, then for n large enough

$$\begin{aligned}
\mathbb{P}(|T_n| \geq \epsilon) \leq & \epsilon^{-2} n^{-2} \sum_{i=k_n^2+1}^{n} \mathbb{E}Z_i^2 \\
\leq & \epsilon^{-2} k_n^{-4} \sum_{i=k_n^2+1}^{n} \mathbb{E}Z_i^2 \\
\leq & \epsilon^{-2} k_n^{-4} \left(n - k_n^2\right) \left(\frac{1}{n-k_n^2} \sum_{i=k_n^2+1}^{n} \mathbb{E}Z_i^2 \right) \\
\leq & \epsilon^{-2} k_n^{-4} \left(n - k_n^2\right)^{1+\rho} \\
\leq & C\epsilon^{-2} k_n^{-4} \left(2k_n + 1\right)^{1+\rho} \\
\leq & C\epsilon^{-2} k_n^{-3+\rho}
\end{aligned}$$

where C is some fixed constant. Then we have

$$\begin{aligned}
\sum_{n=1}^{\infty} \mathbb{P}(|T_n| \geq \epsilon) \leq & \sum_{n=1}^{\infty} C\epsilon^{-2} k_n^{-3+\rho} \\
\leq & \sum_{n=1}^{\infty} C\epsilon^{-2} (\sqrt{n} - 1)^{-3+\rho} \\
\leq & \sum_{n=1}^{4} C\epsilon^{-2} (\sqrt{n} - 1)^{-3+\rho} + 2C\epsilon^{-2} \sum_{n=4}^{\infty} n^{-(3-\rho)/2} \\
< & \infty,
\end{aligned}$$

provided we choose $0 < \rho < 1$. Therefore, by choosing $0 < \rho < 1/2$, it follows from Borel-Cantelli lemma that $T_n \stackrel{a.s.}{\to} 0$ and further $A_n \stackrel{a.s.}{\to} 0$.

B_n converges to 0 almost surely

First of all, we will show $\sigma > 0$ by which we can use Gaussian smoothing to show $B_n \stackrel{a.s.}{\to} 0$.

Lemma B.3. *For $\ell \geq 1$, if $\Sigma(z^\ell, z^\ell) > 0$, then $\Sigma(z^{\ell+1}, z^{\ell+1}) > 0$.*

Proof. We prove by contradiction. Assume $\Sigma(z^{\ell+1}, z^{\ell+1}) = 0$. Then we have

$$0 = \Sigma(z^{\ell+1}, z^{\ell+1}) = \mathbb{E}\phi(z^\ell + z^1)^2 = \mathbb{E}\phi(u^\ell)^2,$$

where $u^\ell \sim \mathcal{N}(0, \Sigma(z^\ell, z^\ell) + \|x\|^2/n_{in})$. It implies $\phi(z) = 0$ almost surely, but it contradicts ϕ is non-constant function since $\Sigma(z^\ell, z^\ell) + \|x\|^2/n_{in} > 0$. $\square$

It follows from Lemma B.3 that $\sigma > 0$. Then $\sigma_n \stackrel{a.s.}{\to} \sigma$, we have $\sigma_n \geq \sigma/2 > 0$ eventually, almost surely. To use Gaussian smoothing, we define following functions

$$f_k(x) := \Phi(g_k^1, \cdots, g_k^\ell, x), \quad F_k(\sigma) := \mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)} f_k(z).$$

By using Gaussian smoothing, we have for large enough n

$$\begin{aligned}
|B_n| \leq& \frac{1}{n}\sum_{k=1}^{n} |F_k(\sigma_n) - F_k(\sigma)| \\
\leq& \frac{1}{n}\sum_{k=1}^{n} \int_{\sigma}^{\sigma_n} |F_k'(t)|\, dt, \quad \text{assume } \sigma \leq \sigma_n \\
\leq& \frac{1}{n}\sum_{k=1}^{n} \int_{\sigma}^{\sigma_n} t^{-1}\mathbb{E}_{z\sim\mathcal{N}(0,1)} \left|f_k(tz)(t^2-1)\right| dt, \quad (a) \\
\leq& \frac{1}{n}\sum_{k=1}^{n} \int_{\sigma}^{\sigma_n} t^{-1}\mathbb{E}_{z\sim\mathcal{N}(0,1)} C_1 e^{C_2\sum_{i=1}^{\ell}|g_k^i| + C_2 t|z| + t} dt, \quad (b) \\
\leq& \frac{1}{n}\sum_{k=1}^{n} \int_{\sigma}^{\sigma_n} t^{-1} C_1 e^{C_2\sum_{i=1}^{\ell}|g_k^i| + C_2 t^2/2 + t} dt, \quad (c) \\
=& C_1\left(\frac{1}{n}\sum_{k=1}^{n} e^{C_2\sum_{i=1}^{\ell}|g_k^i|}\right)(\alpha(\sigma_n) - \alpha(\sigma))
\end{aligned}$$

where (a) is by Lemma A.2 and Jensen's inequality, (b) is because f_k is controllable since Φ is, (c) is by Lemma A.6, and $\alpha(t)$ is the anti-derivative of the function $\dot{\alpha}(t) = t^{-1}C_1 e^{C_2 t^2/2 + t}$. Here, $\dot{\alpha}(t)$ is continuous, so that $\alpha(t)$ is well-defined and continuous. Since $e^{C_2\sum_{i=1}^{\ell}|g_k^i|}$ is controllable, it follows from result for the basic case that

$$\frac{1}{n}\sum_{k=1}^{n} e^{C_2\sum_{i=1}^{\ell}|g_k^i|} \xrightarrow{a.s.} \mathbb{E}_{z\sim\mathcal{N}(0,\Sigma|g^1)} e^{C_2\sum_{i=1}^{\ell}|z^i|}.$$

Since $\sigma_n \overset{a.s.}{\to} \sigma$ and α is continuous, it follows from Lemma A.1 that $\alpha(\sigma_n) \overset{a.s.}{\to} \alpha(\sigma)$ and further

$$|B_n| \leq C_1\left(\frac{1}{n}\sum_{k=1}^{n} e^{C_2\sum_{i=1}^{\ell}|g_k^i|}\right)(\alpha(\sigma_n) - \alpha(\sigma)) \xrightarrow{a.s.} 0.$$

C_n converges to 0 almost surely

Define function $\hat{\Phi}(z^1, \cdots, z^\ell) := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(z^1, \cdots, z^\ell, \sigma z)$. Since Φ is controllable, $\hat{\Phi}$ is also a controllable function. Then it follows from the inductive hypothesis that

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{z\sim\mathcal{N}(0,\sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) &= \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(g_k^1, \cdots, g_k^\ell, \sigma z) \\
&= \frac{1}{n}\sum_{k=1}^{n}\hat{\Phi}(g_k^1, \cdots, g_k^\ell) \\
&\overset{a.s.}{\to} \mathbb{E}\left[\hat{\Phi}(z^1, \ldots, z^\ell)\right] \\
&= \mathbb{E}\left[\mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(z^1, \cdots, z^\ell, \sigma z)\right] \\
&= \mathbb{E}\left[\Phi(z^1, \cdots, z^\ell, z^{\ell+1})\right]
\end{aligned}$$

Thus, $C_n \overset{a.s.}{\to} 0$.

B.4 Proof of Lemma B.1: general case for shared matrices

Now in this section, we prove the desired result when the weight matrices are shared, *i.e.*, $W^\ell = W$. Assume the result holds for ℓ, then we will show the desired result still holds for $\ell + 1$. Note that

$$g^{\ell+1} = Wh^\ell.$$

As W is used before, we have

$$g^i = W h^{i-1}, \quad \forall i \in [\ell].$$

Then define

$$G := \begin{bmatrix} g^1 & g^2 & \cdots & g^\ell \end{bmatrix} \in \mathbb{R}^{n\times \ell}, \quad H := \begin{bmatrix} h^0 & h^1 & \cdots & h^{\ell-1} \end{bmatrix} \in \mathbb{R}^{n\times \ell}. \tag{30}$$

Then we have $G = WH$. Let $\mathcal{B}$ be the σ-algebra spanned by all previous $g^1, g^2, \cdots, g^\ell$. To obtain the conditional distribution of $g^{\ell+1}$ on $\mathcal{B}$, we first compute the conditional distribution of W on $\mathcal{B}$. It follows from Lemma A.3 that

$$\begin{aligned} W|\mathcal{B} =& G\left(H^T H\right)^\dagger H^T + \tilde{W}\Pi_H^T \\ \sim& \mathcal{MN}\left(G(H^T H)^\dagger H^T, I_n, \Pi_H \Pi_H^T/n\right) \end{aligned}$$

where $\Pi = I_n - HH^\dagger$ is the orthogonal projection onto null(H^T), respectively. Therefore, we obtain

$$g^{\ell+1}|\mathcal{B} \sim \mathcal{N}\left(G\left(H^T H\right)^\dagger H^T h^\ell, \|\Pi^T h^\ell\|^2/n I_n\right)$$

or equivalently

$$g_k^{\ell+1}|\mathcal{B} \overset{\text{independent}}{\sim} \mathcal{N}\left(G_k\left(H^T H\right)^\dagger H^T h^\ell, \|\Pi^T h^\ell\|^2/n\right),$$

where $G_k \in \mathbb{R}^{1\times \ell}$ is the k-th row of G.

Since the activation function ϕ is controllable by Lemma A.5, it follows from the inductive hypothesis that

$$(h^i)^T(h^j)/n = \frac{1}{n}\sum_{k=1}^n \phi(g_k^i + g_k^1)\phi(g_k^j + g_k^1) \xrightarrow{a.s.} \mathbb{E}\phi(z^i + z^1)\phi(z^j + z^1) = \Sigma(z^{i+1}, z^{j+1}) \quad \forall i, j.$$

Then we have as $n \to \infty$

$$\begin{aligned} H^T H/n &\xrightarrow{a.s.} \Sigma(Z^\ell, Z^\ell) \\ H^T h^\ell/n &\xrightarrow{a.s.} \Sigma(Z^\ell, z^{\ell+1}) \end{aligned}$$

where $Z^\ell = [z^1 \ \cdots \ z^\ell]^T \in \mathbb{R}^{\ell\times 1}$. Since (pseudo-)inverse is continuous function, we further obtain

$$v_n := \left(H^T H\right)^\dagger H^T h^\ell = \left(H^T H/n\right)^\dagger H^T h^\ell/n \xrightarrow{a.s.} \Sigma(Z^\ell, Z^\ell)^\dagger \Sigma(Z^\ell, z^{\ell+1}) := v. \tag{31}$$

By using the equality $HH^\dagger = H(H^T H)^\dagger H^T$, we have

$$\begin{aligned} \|\Pi^T h^\ell\|^2/n =& \frac{1}{n}(h^\ell)^T\left(I_n - HH^\dagger\right)^2 h^\ell \\ =& \frac{1}{n}(h^\ell)^T\left(I_n - HH^\dagger\right) h^\ell \\ =& \frac{1}{n}(h^\ell)^T h^\ell - \left((h^\ell)^T H/n\right)(H^T H/n)^\dagger \left(H^T h^\ell/n\right) \\ \xrightarrow{a.s.}& \Sigma(z^{\ell+1}, z^{\ell+1}) - \Sigma(z^{\ell+1}, Z^\ell)\Sigma(Z^\ell, Z^\ell)^\dagger \Sigma(Z^\ell, z^{\ell+1}). \end{aligned}$$

By using triangular inequality, we have

$$\left|\frac{1}{n}\sum_{k=1}^n \Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \mathbb{E}\left[\Phi(z^1, \cdots, , z^{\ell+1})\right]\right| \le |A_n| + |B_n| + |C_n| + |D_n|,$$

where

$$A_n = \frac{1}{n}\sum_{k=1}^n \Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n}, \sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) \tag{32}$$

$$B_n = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n}, \sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) - \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n}, \sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) \tag{33}$$

$$C_n = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n}, \sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) - \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_k, \sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) \tag{34}$$

$$D_n = \frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_k, \sigma^2)}\Phi(g_k^1, \cdots, g_k^\ell, z) - \mathbb{E}\left[\Phi(z^1, \cdots, z^{\ell+1})\right] \tag{35}$$

where

$$\mu_{k,n} = G_k^\ell (H^T H)^\dagger H^T h_\ell = G_k^\ell v_n, \tag{36}$$
$$\mu_k = G_k^\ell \Sigma(Z^\ell, Z^\ell)^\dagger \Sigma(Z^\ell, z^{\ell+1}) = G_k^\ell v, \tag{37}$$
$$\sigma_n^2 = \|\Pi^T h^\ell\|^2 \tag{38}$$
$$\sigma^2 = \Sigma(z^{\ell+1}, z^{\ell+1}) - \Sigma(z^{\ell+1}, Z^\ell)\Sigma(Z^\ell, Z^\ell)^\dagger \Sigma(Z^\ell, z^{\ell+1}). \tag{39}$$

B.4.1 A_n converges to 0 almost surely

Define random variables $Z_k = \Phi(g_k^1, \cdots, g_k^\ell, g_k^{\ell+1}) - \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n},\sigma_n^2)}\Phi(g_k^1, \cdots, g_k^\ell, z)$. As $X_k|\mathcal{B}$ are independent, we can easily show X_k are centered and uncorrelated. By using Jensen's inequality, $Z_k^2|\mathcal{B}$ can be upper bounded as follows

$$\mathbb{E}\left[Z_k^2|\mathcal{B}\right] \le 8\mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n},\sigma_n^2)} \left|\Phi(g_k^1, \cdots, g_k^\ell, z)\right|^2 \le 8C_1 e^{2C_2 \sum_{i=1}^\ell |g_k^i|} e^{2C_2|\mu_{k,n}|} e^{2C_2^2\sigma_n^2} \tag{40}$$

As $v_n \overset{a.s.}{\to} v$ by equation (31), we have $\|v_n\| \le 1 + \|v\|$, eventually, almost surely. Thus, for large enough n, we have

$$|\mu_{k,n}| = \left|G_k^\ell (H^T H)^\dagger (H^T h^\ell)\right| = \left|\sum_{i=1}^\ell v_{n,i} g_k^i\right| \le (\|v\|+1)\sum_{i=1}^\ell \left|g_k^i\right|, \tag{41}$$

where we also use the Cauchy-Schwartz inequality and square root inequality. It follows from equation (40) that

$$\mathbb{E}\left[Z_k^2|\mathcal{B}\right] \le 8C_1 e^{(2C_2+\|v\|+1)\sum_{i=1}^\ell |g_k^i|} e^{2C_2^2\sigma_n^2} = \hat{\Phi}(g_k^1, \cdots, g_k^\ell)\cdot e^{2C_2^2\sigma_n^2},$$

where $\hat{\Phi}(x^1, \cdots, x^\ell) := 8C_1 e^{(2C_2+\|v\|+1)\sum_{i=1}^\ell |x_i|}$ is clearly a controllable function. It follows from inductive hypothesis and some basic properties of almost surely convergence in Lemma A.1 that

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[Z_k^2|\mathcal{B}\right] \le \frac{1}{n}\sum_{k=1}^n \hat{\Phi}(g_k^1, \cdots, g_k^\ell)\cdot e^{2C_2^2\sigma_n^2} \xrightarrow{a.s.} \mathbb{E}\left[\hat{\Phi}(z^1, \cdots, z^\ell)\right]\cdot e^{2C_2^2\sigma^2}.$$

As RHS is a deterministic constant, we have $\frac{1}{n}\sum_{k=1}^n \mathbb{E}\left[X_k^2|\mathcal{B}\right] \in o(n^\rho)$ for all $\rho > 0$. Then by using the same argument provided in Section B.3, we have $A_n \overset{a.s.}{\to} 0$.

B_n converges to 0 almost surely

If $\sigma > 0$

In this subsection, we assume $\sigma > 0$. In addition, since $\sigma_n \overset{a.s.}{\to} \sigma$, we have $\sigma_n \ge \sigma/2 > 0$ almost surely for large enough n.

We can obtain the desired result $B_n \overset{a.s.}{\to} 0$ by applying the same argument in Section B.3 to functions f_k and F_k redefined as follows

$$f_k(x) := \Phi(g_k^1, \cdots, g_k^\ell, x), \quad F_k(\sigma) := \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n},\sigma^2)} f_k(z).$$

By using Gaussian smoothing, for large enough n, we have

$$\begin{aligned}
|B_n| \leq & \frac{1}{n}\sum_{k=1}^{n}|F_k(\sigma_n)-F_k(\sigma)| \\
\leq & \frac{1}{n}\sum_{k=1}^{n}\int_{\sigma}^{\sigma_n}|F_k'(t)|\,dt, \quad \text{assume } \sigma \leq \sigma_n \\
\leq & \frac{1}{n}\sum_{k=1}^{n}\int_{\sigma}^{\sigma_n} t^{-1}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left|f_k(\mu_{k,n}+tz)(t^2-1)\right|dt, \quad (a) \\
\leq & \frac{1}{n}\sum_{k=1}^{n}\int_{\sigma}^{\sigma_n} t^{-1}\mathbb{E}_{z\sim\mathcal{N}(0,1)}C_1 e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|+C_2 t|z|+t}dt, \quad (b) \\
\leq & \frac{1}{n}\sum_{k=1}^{n}\int_{\sigma}^{\sigma_n} t^{-1}C_1 e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|+C_2 t^2/2+t}dt, \quad (c) \\
= & C_1\left(\frac{1}{n}\sum_{k=1}^{n}e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|}\right)(\alpha(\sigma_n)-\alpha(\sigma)),
\end{aligned}$$

where (a) is by Lemma A.2, (b) is because f_k is controllable since Φ is, (c) is by Lemma A.6 and equation (41), and $\alpha(t)$ is the anti-derivative of the function $\dot{\alpha}(t) = t^{-1}C_1 e^{C_2 t^2/2+t}$. Here, $\dot{\alpha}(t)$ is continuous, so that $\alpha(t)$ is well-defined and continuous. Since $e^{C\sum_{i=1}^{\ell}|g_k^i|}$ is controllable for any constant C, it follows from the inductive hypothesis that

$$\frac{1}{n}\sum_{k=1}^{n}e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|} \xrightarrow{a.s.} \mathbb{E}\left[e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|z_i|}\right] < \infty.$$

Since $\sigma_n \stackrel{a.s.}{\to} \sigma$ and α is continuous, it follows from Lemma A.1 that $\alpha(\sigma_n) \stackrel{a.s.}{\to} \alpha(\sigma)$ and further

$$|B_n| \leq C_1\left(\frac{1}{n}\sum_{k=1}^{n}e^{(C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|}\right)(\alpha(\sigma_n)-\alpha(\sigma)) \xrightarrow{a.s.} 0.$$

If $\sigma = 0$

In this subsection, we consider when $\sigma = 0$. Note that the argument in the case $\sigma > 0$ also holds if $\sigma = 0$ and $\sigma_n \neq 0$ (infinitely often), because the derivatives $F_k'(t)$ are well-defined if either $\sigma > 0$ or $\sigma_n > 0$. Thus, we only need to analyze the case where $\sigma = 0$ and $\sigma_n = 0$ eventually.

For $\sigma = 0$, we have $\Sigma(z^{\ell+1}, z^{\ell+1}) = \Sigma(z^{\ell+1}, Z^{\ell})\Sigma(Z^{\ell}, Z^{\ell})^{\dagger}\Sigma(Z^{\ell}, z^{\ell+1})$. By Lemma A.4, we have

$$z^{\ell+1} = \Sigma(z^{\ell+1}, Z^{\ell})\Sigma(Z^{\ell}, Z^{\ell})^{\dagger}Z^{\ell} = vZ^{\ell}, \quad a.s.$$

For controllable Φ, we can show the function $\hat{\Phi} : (g_k^1, \cdots, g_k^{\ell}) \mapsto \Phi(g_k^1, \cdots, g_k^{\ell}, G_k^{\ell}v_n)$ is also controllable as follows

$$\begin{aligned}
\left|\hat{\Phi}(g_k^1, \cdots, g_k^{\ell})\right| = & \left|\Phi(g_k^1, \cdots, g_k^{\ell}, G_k^{\ell}v_n)\right| \\
\leq & C_1 e^{C_2\sum_{i=1}^{\ell}|g_k^i|+C_2\left|\sum_{i=1}^{\ell}v_{n,i}g_k^i\right|} \\
\leq & C_1 e^{(2C_2+\|v\|+1)\sum_{i=1}^{\ell}|g_k^i|},
\end{aligned}$$

where the second inequality follows from equation (31). By using the inductive hypothesis, we obtain

$$\begin{aligned}
\frac{1}{n}\sum_{k=1}^{n}\Phi(g_k^1, \cdots, g_k^{\ell}, G_k^{\ell}v_n) = & \frac{1}{n}\sum_{k=1}^{n}\hat{\Phi}(g_k^1, \cdots, g_k^{\ell}) \\
\stackrel{a.s.}{\to} & \mathbb{E}\left[\hat{\Phi}(z^1, \cdots, z^{\ell})\right] \\
= & \mathbb{E}\left[\Phi(z^1, \cdots, z^{\ell}, vZ^{\ell})\right] \\
= & \mathbb{E}\left[\Phi(z^1, \cdots, z^{\ell+1})\right].
\end{aligned} \tag{42}$$

Moreover, as we assume $\sigma_n = 0$ for all large enough n, we obtain $g_k^{\ell+1}|\mathcal{B} = G_k^\ell v_n$ almost surely. Then for large enough n, we obtain

$$\frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_{k,n},\sigma_n^2)}\Phi(G_k^\ell, z) = \frac{1}{n}\sum_{k=1}^n \Phi(g_k^1,\cdots,g_k^\ell,\mu_{k,n}) = \frac{1}{n}\sum_{k=1}^n \Phi(g_k^1,\cdots,g_k^\ell, G_k^\ell v_n) \quad (43)$$

Combining $A_n \overset{a.s.}{\to} 0$ with equations (42) and (43) yields $B_n \overset{a.s.}{\to} 0$.

B.4.2 C_n converges to 0 almost surely

As discussed in Section B.4.1, we can assume $\sigma > 0$. By using Gaussian smoothing again, we can easiliy show $C_n \overset{a.s.}{\to} 0$ since $\mu_{k,n} \overset{a.s.}{\to} \mu_k$. Define functions

$$f_k(x) = \Phi(g_k^1,\cdots,g_k^\ell,x), \quad F_k(\mu) = \mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)} f_k(z).$$

It follows from Lemma A.2 that

$$\begin{aligned}
|C_n| &\le \frac{1}{n}\sum_{k=1}^n |F_k(\mu_{k,n}) - F_k(\mu)| \\
&\le \frac{1}{n}\sum_{k=1}^n \int_{\mu_k}^{\mu_{k,n}} |F_k'(t)|\,dt, \quad \text{assume } \mu_k \le \mu_{k,n} \\
&\le \frac{1}{n}\sum_{k=1}^n \int_{\mu_k}^{\mu_{k,n}} \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\,|f_k(t+\sigma z)|\,|z|\,dt \\
&\le \frac{1}{n}\sum_{k=1}^n \int_{\mu_k}^{\mu_{k,n}} \frac{1}{\sigma}\mathbb{E}_{z\sim\mathcal{N}(0,1)} C_1 e^{C_2\sum_{i=1}^\ell |g_k^i| + C_2 t + (C_2\sigma+1)|z|} dt \\
&\le \frac{1}{\sigma} C_1 e^{(C_2\sigma+1)^2/2}\cdot \frac{1}{n}\sum_{k=1}^n e^{C_2\sum_{i=1}^\ell |g_k^i|}\cdot\left[\beta(\mu_{k,n}) - \beta(\mu_k)\right],
\end{aligned}$$

where $\beta(\mu)$ is the anti-derivative of the function $\dot\beta(t) = e^{C_2 t}$. Here β is well-defined and continuous since $\dot\beta$ is continuous. As $\mu_{k,n} \overset{a.s.}{\to} \mu_k$, it follows from inductive hypothesis and Lemma A.1 that $C_n \overset{a.s.}{\to} 0$.

D_n converges to 0 almost surely

In this section, we can show $D_n \overset{a.s.}{\to} 0$ straightforward from the induction. Define functions

$$\hat\Phi(z^1,\cdots,z^\ell) := \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[\Phi(z^1,\cdots,z^\ell,\sum_{i=1}^\ell v_i z_i + \sigma z)\right].$$

Here $\hat\Phi$ is controllable as Φ is. By applying the inductive hypothesis on $\hat\Phi$, we obtain

$$\begin{aligned}
D_n &= \left|\frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(\mu_k,\sigma^2)}\Phi(g_k^1,\cdots,g_k^\ell,z) - \mathbb{E}\left[\Phi(z^1,\cdots,z^{\ell+1})\right]\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(g_k^1,\cdots,g_k^\ell,\mu_k+\sigma z) - \mathbb{E}_{z^1,\cdots,z^\ell}\mathbb{E}_{z^{\ell+1}|z^1,\cdots,z^\ell}\Phi(z^1,\cdots,z^{\ell+1})\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^n \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(g_k^1,\cdots,g_k^\ell,\mu_k+\sigma z) - \mathbb{E}_{z^1,\cdots,z^\ell}\mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(z^1,\cdots,\sigma z)\right| \\
&= \left|\frac{1}{n}\sum_{k=1}^n \hat\Phi(g_k^1,\cdots,g_k^\ell) - \mathbb{E}_{z^1,\cdots,z^\ell}\hat\Phi(z^1,\cdots,z^\ell)\right| \\
&\overset{a.s.}{\to} 0,
\end{aligned}$$

where we use the fact $\mu_k = G_k^\ell v = \sum_{i=1}^\ell v_i g_k^i$.

C Proof of Corollary 4.2

Define Gaussian random variables $u^\ell(x)$ that is encoded by input x as follows for all $\ell = [2, L-1]$

$$u^1(x) = z^1(x) \tag{44}$$

$$u^\ell(x) = z^\ell(x) + z^1(x). \tag{45}$$

Then we can easily compute the corresponding covariance as follows for $\ell \geq 2$

$$\begin{aligned}
\mathrm{cov}(u^1(x), u^1(x')) &= \mathrm{cov}(z^1(x), z^1(x')) \\
&= \Sigma^1(x, x') \\
\mathrm{cov}(u^\ell(x), u^\ell(x')) &= \mathrm{cov}(z^\ell(x) + z^1(x), z^\ell(x') + z^1(x')) \\
&= \mathrm{cov}(z^\ell(x), z^\ell(x')) + \mathrm{cov}(z^1(x), z^1(x')) \\
&= \Sigma^\ell(x, x') + \Sigma^1(x, x')
\end{aligned}$$

D Proof of Theorem 4.3

This section is deducted to prove the strict positive definiteness of Σ^L. We will prove it by using the notion of *dual activation* and *Hermitian expansion*.

Let $x \sim \mathcal{N}(0,1)$ and $f : \mathbb{R} \to \mathbb{R}$. Then we can define an inner product

$$\langle f, g \rangle := \mathbb{E}_{x \sim \mathcal{N}(0,1)} f(x) g(x).$$

Thus, we define a Hilbert space of functions $\mathcal{H}$, that is, $f \in \mathcal{H}$ if and only if

$$\|f\|^2 = \mathbb{E}_{x \sim \mathcal{N}(0,1)} |f(x)|^2 < \infty.$$

Next, consider the function sequence $1, x, x^2, \cdots$. Clearly, they are independent. Then apply Gram-Schmidt process to the function sequence w.r.t. the inner product we define before, and we obtain $\{h_n\}$ the **(normalized) Hermite polynomial** that is an **orthonormal basis** to the Hilbert space $\mathcal{H}$.

Now, we are ready to introduce *dual activation*. The **dual activation** $\hat{\phi} : [-1,1] \to \mathbb{R}$ of an activation $\phi : \mathbb{R} \to \mathbb{R}$ is defined by

$$\hat{\phi}(\rho) := \mathbb{E}_{(X,Y) \sim \mathcal{N}_\rho} \phi(X) \phi(Y). \tag{46}$$

where $\mathcal{N}_\rho$ is multidimensional Gaussian distribution with mean 0 and covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

Then the **dual kernel** k_ϕ is given by

$$k_\phi(x, x') := \hat{\phi}(\langle x, x' \rangle).$$

If a function $\phi \in \mathcal{H}$, we not only can obtain an expansion by using the orthonormal basis of Hermitian polynomials but also an expansion to the dual activation $\hat{\phi}$ by using the same Hermitian coefficients. As a consequence, the corresponding dual kernel k_ϕ can be shown to be strict positive definite by using the Hermitian expansion.

Lemma D.1. *[11, Lemma 12] If $\phi \in \mathcal{H}$, then*

$$\phi(x) = \sum_{n=0}^{\infty} a_n h_n(x), \tag{47}$$

$$\hat{\phi}(\rho) = \sum_{n=0}^{\infty} a_n^2 \rho^n. \tag{48}$$

where $a_n := \langle h_n, \phi \rangle$ is the ***Hermite coefficients****, and the above is* ***Hermitian expansion****.*

Theorem D.1. *[21, Theorem 3][15, Theorem 1] For a function $f : [-1,1] \to \mathbb{R}$ with $f = \sum_{n=0}^{\infty} b_n h_n$, the kernel $K_f : S^{n_0-1} \times S^{n_0-1} \to \mathbb{R}$ defined by*

$$K_f(x, x') := f(x^T x')$$

is ***strictly positive define*** *for any $n_0 \geq 1$ if and only if the coefficients $b_n > 0$ for infinitely many even and odd integer n.*

Now we are ready to prove the kernel or covariance function Σ^L is strict positive definite by using Gaussian measure techniques on the existence of positive definiteness.

Lemma D.2. *Suppose ϕ is non-polynomial Lipschitz continuous. If Σ^ℓ is strictly positive, then $\Sigma^{\ell+1}$ is also strictly positive definite.*

Proof. Assume the contrary. Then there exists a finite distinct collection $\{x_i\}_{i=1}^n$ and some constants $\{c_i\}_{i=1}^n$ such that

$$0 = \sum_{i,j=1}^{n} c_i c_j \Sigma^{\ell+1}(x_i, x_j) = \mathbb{E}\left[\sum_{i=1}^{n} c_i \phi(u_i)\right]^2.$$

This indicates $\sum_{i=1}^{n} c_i\phi(u_i) = 0$ almost surely. Note that we have the random variables (u_i, u_j) follows Gaussian distribution given by

$$(u_i, u_j) \sim \mathcal{N}(0, A^{\ell}(x_i, x_j)).$$

WLOG, we can assume $c_1 \neq 0$. Then for some $\phi(u_1) \neq 0$, we choose $u_1 = \cdots = u_n = u_2$. Then

$$c_1\phi(u_1) + (c_2 + \cdots + c_n)\phi(u_1) = 0,$$

indicates $c_1 = -(c_2 + \cdots + c_n)$. Then for any $u \neq u'$, we have

$$c_1\phi(u) + (-c_1)\phi(u') = 0$$

This implies $\phi(u) = \phi(u')$, but it contradicts ϕ is non-constant.

□

Lemma D.3. *Suppose ϕ is non-polynomial Lipschitz continuous. Then Σ^2 is strictly positive definite.*

Proof. For $\ell = 2$, we have

$$\Sigma^2(x, x') = \sigma_2^2 \mathbb{E}_{(u,v)\sim\mathcal{N}(0,A^1(x,x'))}\left[\phi(u)\phi(v)\right],$$

where

$$A^1(x, x') = \begin{bmatrix} 1 & \langle x, x'\rangle \\ \langle x', x\rangle & 1 \end{bmatrix}.$$

Then we have

$$\Sigma^2(x, x') = \sigma_2^2 \hat{\mu}(x^T x')$$

where $\mu(x) := \phi(x\sigma_u)$.

Clearly, μ is Lipschitz continuous since ϕ is. Let the expansion of μ in Hermite polynomials $\{h_n\}_{n=0}^{\infty}$ to be given as $\mu = \sum_{n=0}^{\infty} a_n h_n$. Then we can write $\hat{\mu}$ as $\hat{\mu}(\rho) = \sum_{n=0}^{\infty} a_n^2 \rho^n$. Then we have

$$\Sigma^2(x, x') = \sigma_w^2 \hat{\mu}(x^T x') = \sigma_w^2 \sum_{n=0}^{\infty} a_n^2 (x^T x')^n.$$

Since ϕ is assumed non-polynomials, μ is also non-polynomial, and so there are infinitely many number of nonzero a_n in the expansion. Thus, $b_n := a_n^2 > 0$ for infinitely many even and odd numbers. Since $\sigma_w^2 > 0$, we have Σ^2 is strictly positive definite. □

Then we obtain Σ^L is strict positive definite by combining Lemma D.2 and D.3

E Proof of Lemma 4.1

This section we show the limiting covariance function Σ^* is well defined. As each Σ^L satisfies Cauchy-Schwartz inequality, it suffices to show $\Sigma^*(x,x)$ is well defined, which is given in Lemma E.1.

Lemma E.1. *Choose $\sigma_w > 0$ small for which $\beta := \frac{\sigma_w^2}{2}\mathbb{E}|z|^2|z^2-1| < 1$, where z is standard Gaussian random variable. Then we have for every $x \in \mathbb{S}^{n_{in}-1}$ and $\ell \in [2, L]$*

$$\left|\Sigma^{\ell+1}(x,x) - \Sigma^{\ell}(x,x)\right| \le \beta \left|\Sigma^{\ell}(x,x) - \Sigma^{\ell-1}(x,x)\right|. \tag{49}$$

Therefore, $\Sigma^*(x,x) := \lim_{\ell\to\infty}\Sigma^{\ell}(x,x)$ exists uniquely and

$$0 < \Sigma^*(x,x) \le (1+1/\beta)\Sigma^2(x,x). \tag{50}$$

Proof. Fix x and we denote $\sigma_\ell^2 := \Sigma^\ell(x,x)$ to simplify the notation. Define function $\Phi(\sigma) := \mathbb{E}_{u\sim\mathcal{N}(0,\sigma^2)}\phi(u)^2$

$$\begin{aligned}
\sigma_{\ell+1}^2 - \sigma_\ell^2 =& \sigma_w^2\left(\mathbb{E}_{u^{\ell+1}\sim\mathcal{N}(0,\sigma_\ell^2+\sigma_1^2)}\phi(u^{\ell+1})^2 - \mathbb{E}_{u^{\ell}\sim\mathcal{N}(0,\sigma_{\ell-1}^2+\sigma_1^2)}\phi(u^{\ell})^2\right)\\
=&\sigma_w^2\left(\Phi\left(\sqrt{\sigma_\ell^2+\sigma_1^2}\right) - \Phi\left(\sqrt{\sigma_{\ell-1}^2+\sigma_1^2}\right)\right)\\
=&\sigma_w^2\int_{\sqrt{\sigma_{\ell-1}^2+\sigma_1^2}}^{\sqrt{\sigma_\ell^2+\sigma_1^2}}\Phi'(t)dt\\
=&\sigma_w^2\int_{\sqrt{\sigma_{\ell-1}^2+\sigma_1^2}}^{\sqrt{\sigma_\ell^2+\sigma_1^2}}\frac{1}{t}\mathbb{E}_z\phi(tz)^2(z^2-1)dt\\
\le&\sigma_w^2\int_{\sqrt{\sigma_{\ell-1}^2+\sigma_1^2}}^{\sqrt{\sigma_\ell^2+\sigma_1^2}}\frac{1}{t}\mathbb{E}_z\left|tz\right|^2\left|z^2-1\right|dt\\
=&\sigma_w^2\mathbb{E}_z\left|z\right|^2\left|z^2-1\right|\int_{\sqrt{\sigma_{\ell-1}^2+\sigma_1^2}}^{\sqrt{\sigma_\ell^2+\sigma_1^2}}tdt\\
=&\frac{\sigma_w^2\mathbb{E}_z\left|z\right|^2\left|z^2-1\right|}{2}\left|\sigma_\ell^2-\sigma_{\ell-1}^2\right|\\
=&\beta\left|\sigma_\ell^2-\sigma_{\ell-1}^2\right|,
\end{aligned}$$

where $\beta := \frac{\sigma_w^2\mathbb{E}_z|z|^2|z^2-1|}{2}$. As we choose σ_w small such that $\beta < 1$, then the mapping

$$\sigma_{\ell+1}^2 = \mathbb{E}_{u\sim\mathcal{N}(0,\sigma_\ell^2+\sigma_1^2)}\left[\phi(u)^2\right]$$

is a contraction. Thus, it has unique fixed point σ_* such that

$$\sigma_*^2 = \mathbb{E}_{z\sim\mathcal{N}(0,\sigma_*^2+\sigma_1^2)}\phi(u)^2. \tag{51}$$

In addition, let $\tau_\ell^2 = \sigma_\ell^2+\sigma_1^2$ and $\tau_1^2 = \sigma_1^2$, then we have

$$\left|\tau_{\ell+1}^2-\tau_\ell^2\right| = \left|\sigma_{\ell+1}^2-\sigma_\ell^2\right| \le \beta\left|\sigma_\ell^2-\sigma_{\ell-1}^2\right| = \beta\left|\tau_\ell^2-\tau_{\ell-1}^2\right|.$$

Then we repeat this inequality for ℓ times and obtain

$$\left|\tau_{\ell+1}^2-\tau_\ell^2\right| \le \beta^{\ell-1}\left|\tau_2^2-\tau_1^2\right|.$$

As LHS is $\left|\sigma_{\ell+1}^2-\sigma_\ell^2\right|$ and RHS is σ_2^2, we obtain

$$\left|\sigma_{\ell+1}^2-\sigma_\ell^2\right| \le \beta^{\ell-1}\sigma_2^2.$$

Thus, we have

$$\left|\sigma_{\ell+1}^2-\sigma_2^2\right| \le \sum_{s=2}^{\ell}\left|\sigma_{s+1}^2-\sigma_s^2\right| \le \sum_{s=2}^{\ell}\beta^{s-1}\sigma_2^2 \le \frac{1}{\beta}\sigma_2^2.$$

Therefore, we obtain

$$\sigma_\ell^2 \le \left(1+\frac{1}{\beta}\right)\sigma_2^2 < \infty, \quad \forall \ell \ge 2.$$

Now, suppose $\sigma_* = 0$, then we have equation

$$\begin{aligned} 0 =& \sigma_*^2 \\ =& \mathbb{E}_{u\sim\mathcal{N}(0,\sigma_*^2+\sigma_1^2)}\phi(u)^2 \\ =& \mathbb{E}_{u\sim\mathcal{N}(0,\sigma_1^2)}\phi(u)^2 \\ =& \mathbb{E}_{u\sim\mathcal{N}(0,1)}\phi(u)^2 \end{aligned}$$

where we use the fact $\sigma_1^2 = 1$. The equation above implies $\phi(u) = 0$ almost surely, which is impossible since u follows standard Gaussian and ϕ is nonconstant. □

E.1 $\Sigma^*(x,x) = \Sigma^*(x',x')$

In this subsection, we will first show $\Sigma^\ell(x,x) = \Sigma^\ell(x',x')$ for all x, x'. The desired result is obtained by letting $\ell \to \infty$.

Given x_i and x_j, let $A_{ij}^\ell := \Sigma^\ell(x_i, x_j)$. We prove by induction. For the basic case, we have

$$A_{ii}^1 = \mathbb{E}\left|\sigma(x_i^T z)\right|^2 = \mathbb{E}\left|\sigma(u_j^1)\right|^2 = \mathbb{E}\left|\sigma(u_j^1)\right|^2 = A_{jj}^1,$$

where we use the fact $u_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ due to $\|x_i\|^2 = 1$.

Assume the result holds for $\ell-1$. Then we will show the result for ℓ. Note that

$$\mathrm{Var}(u_i^{\ell-1}) = A_{ii}^{\ell-1} + A_{ii}^1 = A_{jj}^{\ell-1} + A_{jj}^1 = \mathrm{Var}(u_j^{\ell-1}),$$

where the last equality holds follow from the inductive hypothesis. As each $u_i^{\ell-1}$ is a centered Gaussian random variable, equal variance implies equal distribution. Then we obtain

$$A_{ii}^\ell = \mathbb{E}_{u_i^{\ell-1}\sim\mathcal{N}(0,A_{ii}^{\ell-1}+A_{ii}^1)}\left|\sigma(u_i^{\ell-1})\right|^2 = \mathbb{E}_{u_j^{\ell-1}\sim\mathcal{N}(0,A_{jj}^{\ell-1}+A_{jj}^1)}\left|\sigma(u_j^\ell)\right|^2 = A_{jj}^\ell.$$

Then let $\ell \to \infty$ and we obtain the desired result.

F Proof of Lemma 4.2

In Theorem 4.1 and Appendix B, we have shown that for any controllable function Φ, $\frac{1}{n}\Phi(g_k^1, \cdots, g_k^\ell)$ converges almost surely. Here we conduct a stronger result by providing the convergence rates.

Lemma F.1. *Let Φ be a controllable function. Then for any $\ell \geq 1$, quantities $\frac{1}{n}\sum_{k=1}^n \Phi(g_k^1(x), g_k^1(x')g_k^\ell(x), g_k^\ell(x'))$ converges to $\mathbb{E}\left[\Phi(z^1(x), z^1(x'), z^\ell(x), z^\ell(x'))\right]$ a.s. with a rate at least $n^{-1/4}$, i.e.,*

$$\left|\frac{1}{n}\sum_{k=1}^n \Phi(g_k^1(x), g_k^1(x')g_k^\ell(x), g_k^\ell(x')) - \mathbb{E}\left[\Phi(z^1(x), z^1(x'), z^\ell(x), z^\ell(x'))\right]\right| \leq n^{-1/4}, \quad a.s. \tag{52}$$

Intuitively, Lemma F.1 provides a convergence rate of width. Th following Lemma provides a convergence rate for depth.

Lemma F.2. *Choose $\sigma_w > 0$ small for which $\gamma := 2\sqrt{2}\sigma_w < 1$. Then for every $x \in \mathbb{S}^{n_{in}-1}$ and for any k and ℓ, we have $\|h^\ell(x) - h^k(x)\| \leq \frac{\gamma^\ell}{1-\gamma}\|h^1\|$ a.s. Consequently, the equilibrium point $h^*(x)$ is uniquely determined a.s. Additionally, we have $\|h^\ell(x)\| \leq \frac{1-\gamma^\ell}{1-\gamma}\|h^1\|$ a.s.*

Now, combines these two convergence rates, we can show the two limits can be switched. As a result, the DEQ f_θ defined in (1) tends to a Gaussian Process.

F.1 Proof of Lemma 4.2

Let $h_n^\ell(x)$ to denote the post-activation at the ℓ-th layer with width m and input x encoded. Let x and x' in $\mathbb{S}^{d-1}$. Then for any $n \leq m$ and $\ell \leq k$, we have that

$$\begin{aligned}
&\left|\frac{1}{n}\left\langle h_n^\ell(x), h_n^\ell(x')\right\rangle - \frac{1}{m}\left\langle h_m^k(x), h_m^k(x')\right\rangle\right| \\
\leq &\left|\frac{1}{n}\left\langle h_n^\ell(x), h_n^\ell(x')\right\rangle - \frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle\right| + \left|\frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle - \frac{1}{m}\left\langle h_m^k(x), h_m^k(x')\right\rangle\right|.
\end{aligned}$$

In the following, we will bound each term. For the first term, by using Lemma F.2, we have

$$\begin{aligned}
&\left|\frac{1}{n}\left\langle h_n^\ell(x), h_n^\ell(x')\right\rangle - \frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle\right| \\
\leq &\frac{1}{n}\|h_n^\ell(x)\| \cdot \|h_n^\ell(x') - h_m^k(x')\| + \frac{1}{n}\|h_n^\ell(x) - h_n^k(x)\| \cdot \|h_n^k(x')\| \\
\leq &\frac{1}{n} \cdot \frac{1}{1-\gamma}\|h_n^1(x)\| \cdot \frac{\gamma^\ell}{1-\gamma}\|h^1(x')\| + \frac{1}{n} \cdot \frac{1}{1-\gamma}\|h_n^1(x)\| \cdot \frac{\gamma^k}{1-\gamma}\|h_n^1(x')\|
\end{aligned}$$

Combining Theorem A.2 with assumption $\|x\| = 1$, we have $\|Ux\| \leq 2\sigma_u\sqrt{n}/\sqrt{n_{in}}$ a.s. WLOG, we assume $\sigma_u = \sqrt{n_{in}}$, then we have $\|h_n^1(x)\| \leq 2\sqrt{n}$ and so

$$\left|\frac{1}{n}\left\langle h_n^\ell(x), h_n^\ell(x')\right\rangle - \frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle\right| \leq \frac{4}{(1-\gamma)^2}\gamma^\ell. \tag{53}$$

For the second term, we have

$$\left|\frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle - \frac{1}{m}\left\langle h_m^k(x), h_m^k(x')\right\rangle\right| \leq I_n + I_m, \tag{54}$$

where

$$I_n = \left|\frac{1}{n}\left\langle h_n^k(x), h_n^k(x')\right\rangle - \Sigma^k(x, x')\right| \tag{55}$$

By using Lemma F.1, we have

$$I_n \leq n^{-1/4} \tag{56}$$

Similarly, $I_m \leq m^{-1/4}$. Then we can combine these and get

$$\left| \frac{1}{n} \left\langle h_n^\ell(x), h_n^\ell(x') \right\rangle - \frac{1}{m} \left\langle h_m^k(x), h_m^k(x') \right\rangle \right| \leq A\gamma^\ell + Bn^{-1/4}, \tag{57}$$

where $A = 4(1-\gamma)^2$ and $B = 2$. Then letting $m, \ell \to \infty$ sequentially yields

$$\left| \frac{1}{n} \left\langle h_n^\ell(x), h_n^\ell(x') \right\rangle - \Sigma^*(x, x') \right| \leq A\gamma^\ell + Bn^{-1/4},$$

F.2 Proof of Lemma F.1

As we discussed before, Lemma B.1 can be easily extended to Lemma B.2 by using same argument on different inputs x and x'. Similarly, here it suffices to show the desired result for single input x, *i.e.*,

$$\left| \frac{1}{n} \sum_{k=1}^{n} \Phi(g_k^1(x), g_k^\ell(x)) - \mathbb{E}\left[\Phi(z^1(x), z^\ell(x),)\right] \right| \leq n^{-1/4}, \quad a.s. \tag{58}$$

F.2.1 Consider the basic case $\ell = 1$

For $\ell = 1$, we have $g^1 = Ux$ and so

$$g_k^1 \overset{i.i.d.}{\sim} \mathcal{N}(0, \|x\|^2/n_{in}).$$

Let $X_k := \Phi(g_k^1) - \mathbb{E}\Phi(g_k^1)$. Then $\mathbb{E}X_k = 0$ and

$$\mathbb{E}\left|X_k\right|^2 = \mathbb{E}\left|\Phi(g_k^1) - \mathbb{E}\Phi(g_k^1)\right|^2 \leq 8\mathbb{E}\left|\Phi(z^1)\right|^2 \leq 8C\mathbb{E}e^{c|z_1|} < \infty,$$

where we use the fact z^1 and g_k^1 are identically distributed.

It follows from Markov's inequality, we have for any $t > 0$

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} \Phi(g_k^1) - \mathbb{E}\Phi(z^1)\right| > t\right] = \mathbb{P}\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right| > t\right] \leq t^{-2}\mathbb{E}\left|\frac{1}{n}\sum_{k=1}^{n} X_k\right|^2 = t^{-2}n^{-1}\mathbb{E}\left|X_k\right|^2.$$

Therefore, we have $\left|\frac{1}{n}\sum_{k=1}^{n} \Phi(g_k^1) - \mathbb{E}\Phi(z^1)\right| \to 0$ in probability as $n \to \infty$. It follows from Levy's Theorem that this convergence is almost surely because X_k are independent. Additionally, for any $\varepsilon, \delta > 0$, let $t = R(n)\varepsilon$ and let RHS be less than δ. Then we obtain

$$R(n) \geq \delta^{-1/2}\epsilon^{-1}\mathbb{E}\left|X_k\right|^2 n^{-1/2},$$

which indicates the convergence rate is at least $n^{-1/2}$.

F.2.2 The general case ℓ

We can use similar argument from Appendix C to obtain the desired result. Lemma B.2 or Lemma B.1 has been shown weight-tied and weight-untied converges to the same Gaussian process. WLOG, we can just focus on the weight-untied case. Let $\mathcal{B}$ be the σ-algebra spanned by g^1 and g^ℓ, then we have

$$g_k^{\ell+1}|\mathcal{B} \overset{i.i.d.}{\sim} \mathcal{N}(0, \|h^\ell\|^2/n).$$

By using the inductive hypothesis, we have

$$\sigma_{\ell,n}^2 := \|h^\ell\|^2/n \overset{a.s.}{\to} \mathbb{E}\left[\phi(z^\ell + z^1)\right] := \sigma_\ell^2 \tag{59}$$

with convergence rate $n^{-1/4}$, *i.e.*,

$$\left|\sigma_{\ell,n}^2 - \sigma_\ell^2\right| \leq n^{-1/4}, \quad a.s. \tag{60}$$

By using triangle inequality, we have

$$\left|\frac{1}{n}\sum_{k=1}^{n} \Phi(g_k^1, g_k^{\ell+1}) - \mathbb{E}\Phi(z^1, z^{\ell+1})\right| \leq |A_n| + |B_n| + |C_n|,$$

where

$$A_n = \frac{1}{n}\sum_{k=1}^{n}\Phi(g_k^1, g_k^{\ell+1}) - \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Phi(g_k^1, \sigma_{\ell,n} z)$$
$$B_n = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Phi(g_k^1, \sigma_{\ell,n} z) - \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Phi(g_k^1, \sigma_{\ell} z)$$
$$C_n = \frac{1}{n}\sum_{k=1}^{n}\mathbb{E}\Phi(g_k^1, \sigma_{\ell} z) - \mathbb{E}\Phi(z^1, z^{\ell+1})$$

Convergence of A_n

Let $Z_k := \Phi(g_k^1, g_k^{\ell+1}) - \mathbb{E}\Phi(g_k^1, \sigma_{\ell,n} z)$. With the same argument in Appendix B, we have $\mathbb{E}[Z_k|\mathcal{B}] = 0$ and $\mathbb{E}\,|Z_k|\mathcal{B}|^2 \leq 8C_1 e^{2C_2|g_k^1|} e^{2C_2^2\sigma_{\ell,n}^2}$. As $\sigma_{\ell,n} \to \sigma_\ell$, we have

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Z_k|\mathcal{B}]^2 \leq 8C_1\left[\frac{1}{n}\sum_{k=1}^{n} e^{2C_2|g_k^1|}\right] e^{2C_2^2\sigma_{\ell,n}^2} \overset{a.s.}{\to} 8C_1\left[\mathbb{E}e^{2C_2|z^1|}\right] e^{2C_2^2\sigma_\ell^2}.$$

Additionally, it follows from Theorem 4.4 that $\sigma_\ell \to \sigma_*$ as $\ell \to \infty$, we obtain $\sigma_\ell \leq C_3\sigma_*$ for some absolute constant C_3. Then for large enough n, we have

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Z_k|\mathcal{B}]^2 \leq 16C_1\left[\mathbb{E}e^{2C_2|z^1|}\right] e^{4C_2^2C_3^2\sigma_*^2}. \tag{61}$$

As RHS is a deterministic constant, we obtain for large enough n

$$\frac{1}{n}\sum_{k=1}^{n}\mathbb{E}[Z_k|\mathcal{B}]^2 \leq n^\rho, \quad \forall \rho > 0.$$

It is worth to note that we obtain the same result in Appendix B. However, RHS of (61) is independent of ℓ. As a consequence, the inequality (61) holds uniformly over all ℓ. This potentially indicates the limits of depth and width commutes. From here, with almost identical argument in Appendix B, we obtain $A_n \overset{a.s.}{\to} 0$ at rate $n^{-1/4}$ by choosing $\rho = 1/2$.

Convergence of B_n

Similarly, we can use the same argument in Appendix B to get

$$|B_n| \leq C_1\left[\frac{1}{n}\sum_{k=1}^{n} e^{C_2|g_k^1|}\right](\alpha(\sigma_{\ell,n}) - \alpha(\sigma_\ell)).$$

As $\frac{1}{n}\sum_{k=1}^{n} e^{C_2|g_k^1|}$ is a controllable function of g_k^1 and $\sigma_{\ell,n} \overset{a.s.}{\to} \sigma_\ell$, the inductive hypothesis implies $B_n \overset{a.s.}{\to} 0$ at a rate $n^{-1/4}$.

Convergence of C_n

Define function $\hat{\Phi}(x) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\Phi(x, \sigma_\ell z)$. Then C_n becomes

$$C_n = \frac{1}{n}\sum_{k=1}^{n}\hat{\Phi}(g_k^1) - \mathbb{E}\hat{\Phi}(z^1).$$

As $\hat{\Phi}$ is controllable since $\Phi$is, the inductive hypothesis implies directly $C_n \overset{a.s.}{\to} 0$ at a rate of $n^{-1/4}$.

F.3 Proof of Lemma F.2

It follows from Theorem A.2 that $\frac{1}{\sqrt{n}}\|W\| \leq 2\sqrt{2}\sigma_w$ a.s. Then we can choose a small σ_w for which $\gamma := 2\sqrt{2}\sigma_w < 1$. Then for any $\ell \geq 0$, the Lipschitz continuity of ϕ implies

$$\begin{aligned}
\|h^{\ell+1} - h^\ell\| =& \frac{1}{\sqrt{n}}\|\phi(Wh^\ell + g^1) - \phi(Wh^{\ell-1} + g^1)\| \\
\leq& \frac{1}{\sqrt{n}}\|Wh^\ell - Wh^{\ell-1}\| \\
\leq& \frac{1}{\sqrt{n}}\|W\|\|h^\ell - h^{\ell-1}\| \\
\leq& \gamma\|h^\ell - h^{\ell-1}\|.
\end{aligned}$$

Thus, we repeat this argument ℓ times and obtain

$$\|h^{\ell+1} - h^\ell\| \leq \gamma^\ell\|h^1 - h^0\| = \gamma^\ell\|h^1\|$$

From here, for any $k \geq \ell \geq 0$, we have

$$\|h^\ell - h^k\| \leq \sum_{s=\ell}^{k-1}\|h^s - h^{s+1}\| \leq \sum_{s=\ell}^{k-1}\gamma^s\|h^1\| \leq \frac{\gamma^\ell(1-\gamma^{k-\ell})}{1-\gamma}\|h^1\|. \tag{62}$$

Thus, it follows from the completeness of $\mathbb{R}^m$ that the unique $h^*(x)$ exists. Additionally, let $k \to \infty$, then we have

$$\|h^\ell - h^*\| \leq \frac{\gamma^\ell}{1-\gamma}\|h^1\|.$$

Let $\ell = 0$, then we obtain

$$\|h^k\| \leq \frac{1-\gamma^k}{1-\gamma}\|h^1\|.$$

G Proof of Theorem 4.4

By condition on the values of h^*, the outputs

$$f_{\theta,k}(x) = \langle v_k, h^* \rangle$$

are *i.i.d.*centered Gaussian random variables with covariance

$$\hat{\Sigma}(x, x') = \langle h^*(x), h^*(x') \rangle / n.$$

It follows from Lemma 4.2 that

$$\hat{\Sigma}(x, x') \overset{a.s.}{\to} \Sigma^*(x, x').$$

Specifically, the covariance Σ^* is deterministic and hence independent to h^*. Consequently, the conditioned and unconditioned distributions of $f_{\theta,k}$ are equal in the limit of $n \to \infty$: they are *i.i.d.*centered Gaussian random variables with covariance Σ^*.

H Proof of Theorem 4.5

Equipped with the notion of dual activation and Theorem D.1, we are ready to prove Theorem 4.5, *i.e.*, Σ^* is strict positive definite.

By Lemma E.1, we have $\Sigma^*(x,x) = \Sigma^*(x',x') := c$ and $0 < c < \infty$ for all x, x'. Then we have

$$\Sigma^*(x,x') = \mathbb{E}_{u(x),u(x')\sim\mathcal{N}(0,A^*)}\left[\phi(u(x))\phi(u(x'))\right]$$

where

$$A^* = \begin{bmatrix} \Sigma^*(x,x) + \Sigma^1(x,x) & \Sigma^*(x,x') + \Sigma^1(x,x') \\ \Sigma^*(x',x) + \Sigma^1(x',x) & \Sigma^*(x,x) + \Sigma^1(x',x') \end{bmatrix} = \begin{bmatrix} c+1 & \Sigma^*(x,x') + \langle x, x'\rangle \\ \Sigma^*(x,x') + \langle x, x'\rangle & c+1 \end{bmatrix}.$$

By changing variable with $u(x) = \sqrt{c+1}z(x)$, we obtain

$$\Sigma^*(x,x') = \mathbb{E}\left[\mu(z(x))\mu(z(x'))\right] = \hat{\mu}\left(\frac{\Sigma^*(x,x') + \langle x, x'\rangle}{c+1}\right),$$

where $\hat{\mu} : [-1,1] \to \mathbb{R}$ is dual activation of activation function $\mu(z) := \phi(\sqrt{c+1}z)$.

Let $\mu = \sum_{n=0}^{\infty} a_n h_n$ be the Hermite expansion, then we obtain $\hat{\mu}$ as

$$\hat{\mu}(\rho) = \sum_{n=0}^{\infty} a_n^2 \rho^n.$$

Therefore, Σ^* has the expression

$$\Sigma^*(x,x') = \sum_{n=0}^{\infty} a_n^2 \left(\frac{\Sigma^*(x,x') + \langle x, x'\rangle}{c+1}\right)^n.$$

Since ϕ is non-polynomial, so is μ, and hence, there is an infinite number of nonzero a_n's. By Theorem 2, we can conclude that Σ^* is strictly positive definite and complete the proof.

I Additional Experimental Results

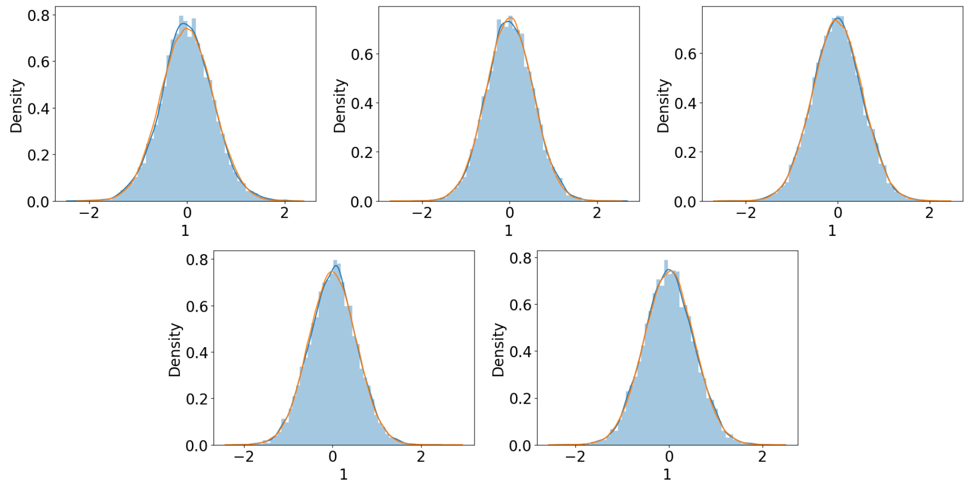

Figure 5: Histplot of the output distributions for five neural networks with widths 10, 50, 100, 200, 1000 (left to right); KS statistics: $0.02641, 0.00677, 0.00550, 0.00321, 0.00302$, pvalue: $9,74 \times 10^{-31}, 0.0202, 0.0969, 0.6808, 0.7498$.