Supplementary Material

A. Empirical Evaluation of Conditions in Hypothesis 5

We consider a sequence of datapoints of increasing m by starting with a compressed low dimensional datapoint and decreasing the amount of compression, evaluating the asymptotic bound $m \max_{i=1}^{m} \left| \frac{x_i^{(m)}}{\|\mathbf{x}^{(m)}\|} \right|^4$ for each m. Figure 5 shows plots of the asymptotic bound for two datasets.

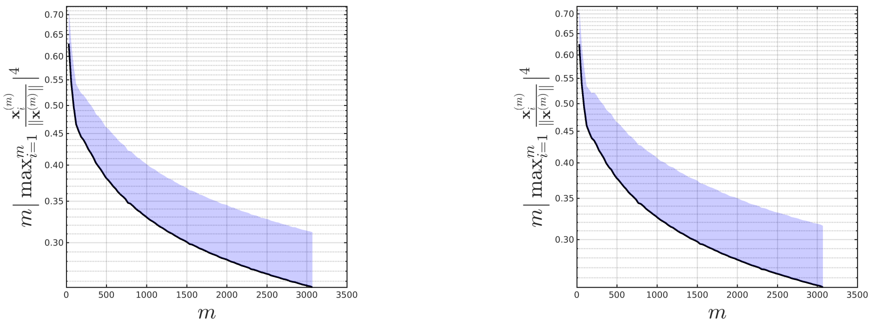

Figure 5. Asymptotic error in the application of the CLT to neural network kernels. The solid line is an average over 1000 randomly sampled datapoints and the shaded region represents 1 standard deviation in the worst-case direction. Data is preprocessed so that each dimension is in the range $[0, 255]$. (Left) CIFAR10 and (Right) CIFAR100 (Krizhevsky & Hinton, 2009). The images are compressed using Bicubic Interpolation.

The plots suggest that Hypothesis 5 makes reasonable assumptions on high dimensional datasets.

B. Proof of Proposition 4

Proof. The LReLU activation function is $\sigma(z) = \big(a + (1-a)\Theta(z)\big)z$. Expanding, we have

$$\begin{aligned} k(\mathbf{x},\mathbf{y}) &= \int_{\mathbb{R}^m} \sigma(\mathbf{w}\cdot\mathbf{x})\sigma(\mathbf{w}\cdot\mathbf{y}) f(\mathbf{w})\,\mathbf{dw}, \\ &= \int_{\mathbb{R}^m} \Big(a^2 + a(1-a)\Theta(\mathbf{w}\cdot\mathbf{y}) + a(1-a)\Theta(\mathbf{w}\cdot\mathbf{x}) + (1-a)^2\Theta(\mathbf{w}\cdot\mathbf{x})\Theta(\mathbf{w}\cdot\mathbf{x})\Big)(\mathbf{w}\cdot\mathbf{x})(\mathbf{w}\cdot\mathbf{y}) f(\mathbf{w})\,\mathbf{dw}. \end{aligned}$$

Using linearity of the integral, we have the superposition of the four integrals $k_1 = a^2\mathbb{E}\big[(\mathbf{W}\cdot\mathbf{x})(\mathbf{W}\cdot\mathbf{y})\big]$, $k_2 = a(1-a)\mathbb{E}\big[\Theta(\mathbf{W}\cdot\mathbf{x})(\mathbf{W}\cdot\mathbf{x})(\mathbf{W}\cdot\mathbf{y})\big]$, $k_3 = a(1-a)\mathbb{E}\big[\Theta(\mathbf{W}\cdot\mathbf{y})(\mathbf{W}\cdot\mathbf{x})(\mathbf{W}\cdot\mathbf{y})\big]$ and $k_4 = (1-a)^2\mathbb{E}\big[\Theta(\mathbf{W}\cdot\mathbf{x})\Theta(\mathbf{W}\cdot\mathbf{y})(\mathbf{W}\cdot\mathbf{x})(\mathbf{W}\cdot\mathbf{y})\big]$.

Now $k_1(\mathbf{x},\mathbf{y}) = a^2\mathbb{E}[W_i^2]\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0$. To see this, rotate the coordinate system as before. Then, either solve the integral directly using the fact that the weights are uncorrelated or differentiate twice and solve the homogeneous IVP with initial conditions $k(0) = a^2\mathbb{E}[W_i^2]\|\mathbf{x}\|\|\mathbf{y}\|$ and $k'(0) = 0$.

After rotating the coordinate system, differentiating $k_2(\mathbf{x},\mathbf{y})$ twice results in a homogeneous IVP with $k(0) = a(1-a)\frac{\mathbb{E}[W_i^2]}{2}\|\mathbf{x}\|\|\mathbf{y}\|$ and $k'(0) = 0$, the solution of which is $k_2(\mathbf{x},\mathbf{y}) = a(1-a)\frac{\mathbb{E}[W_i^2]}{2}\|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0$. Note that by symmetry, $k_2(\mathbf{x},\mathbf{y}) = k_3(\mathbf{x},\mathbf{y})$.

The last remaining integral, $k_4(\mathbf{x},\mathbf{y})$, is just a multiple of the Arc-Cosine kernel. □

C. Other Asymptotic Kernels

Corollary 10 (Asmptotic Kernels: $1-\epsilon$ Exponent-Dominated Activation Functions)**.** *Consider the same scenario as in Corollary 7, with the exception that the activation functions are replaced by some continuous σ such that $|\sigma(z)| \leq M|z|^{1-\epsilon}$ for all $z \in \mathbb{R}$, some $\epsilon > 0$, and some $M \in (0,\infty)$, then for all $s \geq 2$*

$$\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)},\mathbf{y}^{(m)}\big) = k_g^{(s)}\big(\mathbf{x}^{(s)},\mathbf{y}^{(s)}\big) = \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big].$$

Proof. We have $\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = \lim_{m\to\infty} \mathbb{E}\big[\sigma(Z_1^{(m)})\sigma(Z_2^{(m)})\big]$ and we would like to bring the limit inside the expected value. By Theorem 6 and Theorem 25.12 of Billingsley (1995), it suffices to show that $\sigma(Z_1)\sigma(Z_2)$ is uniformly integrable. Define h to be the joint PDF of $\mathbf{Z}$. As in (25.13) of Billingsley (1995), we have

$$\lim_{\alpha\to\infty} \int_{|\sigma(z_1)\sigma(z_2)|>\alpha} |\sigma(z_1)\sigma(z_2)| h(z_1, z_2)\, dz_1 dz_2 \le \lim_{\alpha\to\infty} \frac{1}{\alpha^{\epsilon}} \mathbb{E}\Big[\big|\sigma(Z_1)\sigma(Z_2)\big|^{1+\epsilon}\Big],$$

so it suffices to show that $\mathbb{E}\Big[\big|\sigma(Z_1)\sigma(Z_2)\big|^{1+\epsilon}\Big]$ is bounded. We have

$$\begin{aligned} \mathbb{E}\Big[\big|\sigma(Z_1^{(m)})\sigma(Z_2^{(m)})\big|^{1+\epsilon}\Big] &\le M^2 \mathbb{E}\Big[\big|Z_1^{(m)} Z_2^{(m)}\big|\Big], \\ &\le M^2 \sqrt{\mathbb{E}\Big[\big(Z_1^{(m)}\big)^2\Big]\mathbb{E}\Big[\big(Z_2^{(m)}\big)^2\Big]}, \\ &= M^2 \mathbb{E}[W_i^2] \|\mathbf{x}\| \|\mathbf{y}\| < \infty, \end{aligned}$$

and so

$$\lim_{m\to\infty} k_f^{(m)}(\mathbf{x}, \mathbf{y}) = \mathbb{E}\big[\lim_{m\to\infty} \sigma(Z_1^{(m)})\sigma(Z_2^{(m)})\big] = \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big].$$

□

Corollary 11 (Asymptotic Kernels: Bounded and Continuous Activation Functions)**.** *Consider the same scenario as in Corollary 7, with the exception that the activation functions are replaced by some bounded, continuous σ. Then for all $s \ge 2$*

$$\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = k_g^{(s)}\big(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}\big) = \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big].$$

Proof. This is a direct application of the Portmanteau Lemma to the result in Theorem 6:

$$\begin{aligned} &\Big[\sigma(\mathbf{W}^{(m)} \cdot \mathbf{x}^{(m)})\sigma(\mathbf{W}^{(m)} \cdot \mathbf{y}^{(m)}) \xrightarrow{D} \sigma(Z_1)\sigma(Z_2)\Big] \\ \implies &\Big[\mathbb{E}\big[\sigma(\mathbf{W}^{(m)} \cdot \mathbf{x}^{(m)})\sigma(\mathbf{W}^{(m)} \cdot \mathbf{y}^{(m)})\big] \to \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big]\Big] \end{aligned}$$

for all bounded, continuous σ. □

Corollary 12 (Asymptotic Kernels: LReLU)**.** *Consider the same scenario as in Corollary 7, with the exception that the activation functions are replaced by the Leaky ReLU $\sigma(z) = \Theta(z)z + a\Theta(-z)z, \quad a \in (0,1)$. Then for all $s \ge 2$*

$$\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = k_g^{(s)}\big(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}\big) = \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big].$$

Proof. As before, it suffices to show uniform integrability of the random variable $\sigma\big(Z_1^{(m)}\big)\sigma\big(Z_2^{(m)}\big) = \Theta\big(Z_1^{(m)}\big)\Theta\big(Z_2^{(m)}\big)Z_1^{(m)}Z_2^{(m)} + a\Theta\big(-Z_1^{(m)}\big)\Theta\big(Z_2^{(m)}\big)Z_1^{(m)}Z_2^{(m)} + \Theta\big(Z_1^{(m)}\big)\Theta\big(-Z_2^{(m)}\big)Z_1^{(m)}Z_2^{(m)} + a^2\Theta\big(-Z_1^{(m)}\big)\Theta\big(-Z_2^{(m)}\big)Z_1^{(m)}Z_2^{(m)}$. Each of these terms taken individually is uniformly integrable by the same argument as in Corollary 7. A linear combination of uniformly integrable random variables is uniformly integrable. Thus the random variable is uniformly integrable and as before the result holds. □

Corollary 13 (Asymptotic Kernels: ELU)**.** *Consider the same scenario as in Corollary 7, with the exception that the activation functions are replaced by the ELU $\sigma(z) = \Theta(z)z + \Theta(-z)(e^z - 1)$. Then for all $s \ge 2$*

$$\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = k_g^{(s)}\big(\mathbf{x}^{(s)}, \mathbf{y}^{(s)}\big) = \mathbb{E}\big[\sigma(Z_1)\sigma(Z_2)\big].$$

Proof. By Theorem 6, the random variable $\sigma\big(Z_1^{(m)}\big)\sigma\big(Z_2^{(m)}\big) = \Theta\big(Z_1^{(m)}\big)\Theta\big(Z_2^{(m)}\big)Z_1^{(m)}Z_2^{(m)} + \Theta\big(Z_1^{(m)}\big)Z_1^{(m)}\Theta\big(-Z_2^{(m)}\big)\big(e^{Z_2^{(m)}} - 1\big) + \Theta\big(Z_2^{(m)}\big)Z_2^{(m)}\Theta\big(-Z_1^{(m)}\big)\big(e^{Z_1^{(m)}} - 1\big) + \Theta\big(-Z_1^{(m)}\big)\big(e^{Z_1^{(m)}} - 1\big)\Theta\big(-Z_2^{(m)}\big)\big(e^{Z_2^{(m)}} - 1\big)$ converges in distribution to $\sigma\big(Z_1\big)\sigma\big(Z_2\big)$. Call these terms $T_1^{(m)}, T_2^{(m)}, T_3^{(m)}$, and $T_4^{(m)}$ respectively. Due to linearity of the limit and $\mathbb{E}$,

$$\lim_{m\to\infty} k_f^{(m)}\big(\mathbf{x}^{(m)}, \mathbf{y}^{(m)}\big) = \lim_{m\to\infty} \mathbb{E}[T_1^{(m)}] + \lim_{m\to\infty} \mathbb{E}[T_2^{(m)}] + \lim_{m\to\infty} \mathbb{E}[T_3^{(m)}] + \lim_{m\to\infty} \mathbb{E}[T_4^{(m)}].$$

The first term converges by Corollary 7. The fourth term converges by Corollary 11. The quantity of interest for uniform integrability in the second and third term is the limit as $\alpha \to \infty$

$$\begin{aligned}\Big(\int_{\substack{z_1|e^{z_2}-1|>\alpha\\ z_1>0\\ z_2<0}} z_1|e^{z_2}-1|h(z_1,z_2)\,dz_1dz_2\Big)^2 &= \Big(\mathbb{E}\big[z_1|e^{z_2}-1|\Theta(z_1)\Theta(-z_2)\Theta\big(z_1|e^{z_2}-1|-\alpha\big)\big]\Big)^2,\\ &\le \mathbb{E}\Big[z_1^2\Theta(z_1)\Theta(-z_2)\Theta\big(z_1|e^{z_2}-1|-\alpha\big)\Big]\\ &\quad\mathbb{E}\Big[(e^{z_2}-1)^2\Theta(z_1)\Theta(-z_2)\Theta\big(z_1|e^{z_2}-1|-\alpha\big)\Big],\end{aligned}$$

By the Monotone Convergence Theorem, the first factor evaluates as 0 in the limit using the same argument as in Corollary 7. The second factor is at least bounded by 1 because the argument of $\mathbb{E}$ is always less than 1. So we have uniform integrability, and $\lim_{m\to\infty}\mathbb{E}[T_2^{(m)}]$ and $\lim_{m\to\infty}\mathbb{E}[T_3^{(m)}]$ converge. □

D. Relation to Other Work

From Theorem 6 of the main text, we have that

$$\lim_{m\to\infty} k_f^{(m)}(\theta_0) = \lim_{m\to\infty} \mathbb{E}\big[\sigma(Z_1^{(m)})\sigma(Z_2^{(m)})\big], \quad (Z_1^{(m)}, Z_2^{(m)})^T \xrightarrow{D} N\big(\mathbf{0}, \Sigma\big),$$

with $\Sigma = \mathbb{E}[W_i^2]\begin{bmatrix}\|\mathbf{x}\|^2 & \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0\\ \|\mathbf{x}\|\|\mathbf{y}\|\cos\theta_0 & \|\mathbf{y}\|^2\end{bmatrix}$. If the limit could be moved inside the expectation, the right hand side would resemble the definition of a dual activation, given by Daniely et al. (2016), which follows naturally from the definition of the kernel for the special case of *Gaussian* weights. We have shown that asymptotically for certain activation functions, the limit can indeed be moved inside the expectation and a large class of weight distributions may be treated as Gaussian. Therefore, much of the dual activation results apply to random neural networks operating on high dimensional data from a wide range of distributions.