On Random Kernels of Residual Architectures
Supplementary Material

Etai Littwin[*1] **Tomer Galanti**[*1] **Lior Wolf**[1]

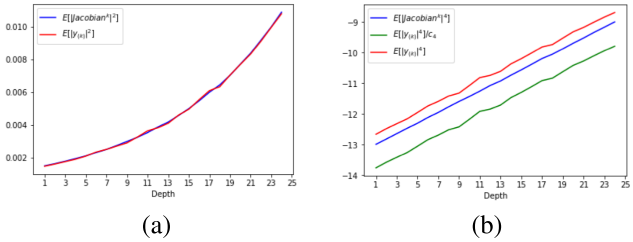
[1]School of Computer Science, Tel Aviv University, Tel Aviv, Israel

1 ABSOLUTE ACCURACY RATES ON UCI DATASETS

In Tab. 1, we report the absolute accuracy rates of kernel regression over random gradient features extracted from the fully connected architectures (e.g., vanilla ReLU networks, ResNets and DenseNets) with three layers and widths 10, 100 and 500 and of kernel regression over the width limit kernel. As can be seen, the various models achieve comparable results on all dataset.

2 VALIDATING THE DUALITY PRINCIPLE

To validate Thm. 2, we estimated the second and fourth moments of the per-layer Jacobian $\|J^{\mathbf{k}}\|_2$ and the squared norm of the output of the corresponding reduced architecture $\|f_{(\mathbf{k})}(x;w)\|_2$ for ResNet architectures (with $m = 2$, $\alpha_l = 0.3$) with a varying number of layers. The results were obtained from the simulated results of 200 independent runs per depth, where the value for k is random for each depth. As can be seen in Fig. 1, the mean of both $\|J^{\mathbf{k}}\|_2^2$ and $\|f_{(\mathbf{k})}(x;w)\|_2^2$ closely match, while the fourth moment $\mathbb{E}[\|J^{\mathbf{k}}\|_2^4]$ is upper and lower bounded by the corresponding moments of the output, as predicted in Thm. 2.



(a) (b)

Figure 1: The second **(a)** and fourth **(b)** moments, in log scale, of the per layer Jacobian norm $\|J^{\mathbf{k}}\|_2$ and the squared norm of the output of the corresponding reduced architecture $\|f_{(\mathbf{k})}(x;w)\|_2$.

[*] Equal contribution

	Vanilla network				ResNet				DenseNet			
Dataset	10	100	500	limit	10	100	500	limit	10	100	500	limit
Abalone	0.51	0.54	0.53	0.54	0.60	0.55	0.55	0.56	0.51	0.52	0.53	0.54
Adult	0.74	0.73	0.76	0.78	0.73	0.72	0.76	0.76	0.76	0.75	0.76	0.74
Bank	0.86	0.85	0.87	0.88	0.87	0.84	0.87	0.88	0.86	0.85	0.88	0.87
Car	0.75	0.81	0.88	0.90	0.76	0.81	0.87	0.89	0.74	0.83	0.88	0.89
Cardiotocography_10clases	0.67	0.73	0.77	0.80	0.70	0.73	0.78	0.78	0.68	0.72	0.77	0.78
Chess_krvk	0.25	0.28	0.36	0.39	0.28	0.29	0.35	0.38	0.28	0.31	0.36	0.38
Chess_krvkp	0.85	0.96	0.97	0.97	0.89	0.96	0.97	0.97	0.87	0.95	0.97	0.97
Connect 4	0.66	0.68	0.73	0.74	0.67	0.68	0.72	0.74	0.68	0.68	0.73	0.74
Contrac	0.46	0.44	0.48	0.48	0.47	0.44	0.49	0.50	0.48	0.43	0.49	0.50
Hill-Valley	0.52	0.57	0.59	0.61	0.53	0.57	0.60	0.57	0.47	0.57	0.63	0.57
Image-Segmentation	0.69	0.75	0.75	0.75	0.72	0.75	0.75	0.75	0.70	0.75	0.75	0.75
Led-Display	0.65	0.68	0.68	0.67	0.65	0.68	0.69	0.65	0.65	0.68	0.68	0.60
Letter	0.56	0.74	0.79	0.80	0.61	0.74	0.80	0.81	0.56	0.73	0.79	0.81
Magic	0.78	0.72	0.80	0.81	0.81	0.73	0.78	0.82	0.80	0.75	0.79	0.82
Molec-biol-splice	0.56	0.74	0.79	0.80	0.62	0.73	0.78	0.77	0.79	0.61	0.68	0.78
Mushroom	0.98	0.99	0.99	0.99	0.99	0.99	0.99	1.0	0.98	1.0	1.0	1.0
Nursery	0.79	0.87	0.92	0.92	0.84	0.86	0.91	0.93	0.80	0.88	0.92	0.93
Oocytes_merluccius_nucleus_4d	0.75	0.74	0.77	0.79	0.78	0.75	0.77	0.77	0.75	0.72	0.76	0.77
Oocytes_merluccius_states_2f	0.87	0.90	0.92	0.92	0.89	0.90	0.92	0.91	0.88	0.90	0.91	0.91
Optical	0.84	0.97	0.98	0.98	0.91	0.97	0.98	0.98	0.87	0.98	0.97	0.98
Ozone	0.94	0.96	0.96	0.96	0.96	0.96	0.97	0.97	0.96	0.96	0.96	0.97
Page_blocks	0.90	0.95	0.95	0.96	0.95	0.95	0.96	0.96	0.95	0.94	0.95	0.96
Pendigits	0.92	0.98	0.98	0.99	0.95	0.98	0.99	0.99	0.93	0.98	0.99	0.99
Plants_margin	0.46	0.68	0.76	0.77	0.54	0.70	0.80	0.77	0.48	0.63	0.74	0.77
Plants_texture	0.57	0.75	0.79	0.80	0.63	0.76	0.80	0.79	0.56	0.73	0.77	0.80
Plants_shape	0.42	0.47	0.52	0.55	0.42	0.50	0.54	0.53	0.38	0.44	0.49	0.53
Ringnorm	0.66	0.66	0.71	0.72	0.69	0.64	0.70	0.73	0.67	0.67	0.71	0.72
Semeion	0.70	0.90	0.93	0.93	0.80	0.92	0.93	0.93	0.72	0.81	0.92	0.92
Spambase	0.82	0.89	0.91	0.92	0.85	0.90	0.91	0.88	0.86	0.89	0.90	0.88
Statlog_german_credit	0.61	0.71	0.73	0.74	0.65	0.70	0.72	0.74	0.64	0.71	0.73	0.74
Statlog_image	0.90	0.93	0.95	0.95	0.91	0.93	0.95	0.95	0.90	0.93	0.95	0.95
Statlog_landsat	0.82	0.84	0.86	0.87	0.83	0.85	0.86	0.87	0.83	0.84	0.86	0.86
Statlog_shuttle	0.97	0.98	0.98	0.98	0.97	0.98	0.98	0.99	0.98	0.98	0.98	0.98
Steel_plates	0.66	0.70	0.74	0.75	0.70	0.70	0.73	0.74	0.66	0.70	0.73	0.75
Thyroid	0.92	0.93	0.95	0.95	0.94	0.94	0.95	0.95	0.94	0.94	0.95	0.95
Titanic	0.54	0.57	0.58	0.70	0.60	0.55	0.61	0.54	0.71	0.67	0.53	0.47
Twonorm	0.90	0.95	0.96	0.96	0.93	0.93	0.96	0.96	0.90	0.95	0.96	0.97
Waveform	0.75	0.74	0.80	0.81	0.77	0.73	0.80	0.81	0.74	0.75	0.81	0.81
Wall_following	0.71	0.79	0.83	0.84	0.73	0.80	0.83	0.83	0.71	0.78	0.82	0.80
Waveform_Noise	0.62	0.77	0.81	0.81	0.70	0.75	0.80	0.82	0.67	0.74	0.81	0.83
Wine_quality_red	0.53	0.55	0.58	0.59	0.54	0.54	0.58	0.60	0.54	0.56	0.59	0.60
Wine_quality_white	0.48	0.47	0.51	0.52	0.48	0.46	0.50	0.52	0.48	0.48	0.51	0.52
Yeast	0.49	0.45	0.50	0.50	0.51	0.45	0.49	0.50	0.50	0.46	0.50	0.51

Table 1: Results of kernel regression over random gradient features on UCI for architectures with 3 layers and widths 10, 100 and 500. The results are compared with the performance of the width limit kernels associated with each architecture.

3 USEFUL LEMMAS

Lemma 1. *Let $f(x;w)$ be a neural network (e.g., vanilla ReLU, ResNet, DenseNet) with N parameters. Let $g(x;w)$ be a pre-activation neuron within $f(x;w)$. Let $x \neq 0$ be an arbitrary input. Then, the set $\{w \mid g(x;w)=0\}$ is of measure zero.*

Proof. We prove the claim by induction on the depth of $g(x;w)$. We denote by $v \in \mathbb{R}^{N_1}$ the subset of w of weights involved in the computation of $g(x;w)$ and by $u \in \mathbb{R}^{N_2}$ the rest of the weights. For simplicity, we will denote $g(x;v) := g(x;w)$.

Base case: Assume $g(x;w)$ is a neuron in the first hidden layer of $f(x;w)$. Then, $g(x;w) = \langle v, x\rangle$, where v is a vector of weights, subset to w. We notice that since $x \neq 0$, the zero set $\{w \mid g(x;w)=0\} = \{u \mid \langle u, x\rangle = 0\} \times \mathbb{R}^{N_2}$ is of dimension $N-1$. Therefore, $\{w \mid g(x;w)=0\}$ is of measure zero.

Induction hypothesis: Assume that for any neuron $g(x;w)$ in the k'th layer, the set $\{w \mid g(x;w)=0\}$ is of measure 0.

Induction step: Let neuron $g(x;w)$ in the $(k+1)$'th layer. Then, we have:

$$g(x;w) = \langle \hat{v}, \hat{g}(x; v \setminus \hat{v})\rangle \tag{1}$$

where $\hat{v}$ are the weights of the specific neuron $g(x;w)$, $\hat{g}(x;v\setminus\hat{v})$ is a concatenation of the neurons that serve as inputs to $g(x;w)$ in the network $f(x;w)$ and $v \setminus \hat{v}$ denotes the set of weights involved in the computation of these neurons.

Let $\hat{g}_1(x;v\setminus\hat{v})$ be the first coordinate of $\hat{g}(x;v\setminus\hat{v})$.

$$\begin{aligned}\{v \mid g(x;v)=0\} \subset &\{v \mid \hat{g}_1(x;v\setminus\hat{v}) \neq 0, g(x;v)=0\} \cup \{v \mid \hat{g}_1(x;v\setminus\hat{v}) = 0, g(x;v)=0\} \\ \subset &\{v \mid \hat{g}_1(x;v\setminus\hat{v}) \neq 0, g(x;v)=0\} \cup \mathbb{R} \times \{v \setminus \hat{v}_1 \mid \hat{g}_1(x;v\setminus\hat{v}) = 0\}\end{aligned} \tag{2}$$

We would like to prove that each set in this union is of measure zero. This will conclude the proof, since a union of measure zero sets is measure zero as well. We note that by the induction hypothesis, the set $\{v \setminus \hat{v}_1 \mid \hat{g}_1(x;v\setminus\hat{v}) = 0\}$ is of measure zero. In particular, $\mathbb{R} \times \{v \setminus \hat{v}_1 \mid \hat{g}_1(x;v\setminus\hat{v}) = 0\}$ is of measure zero. On the other hand, for any $v \setminus \hat{v}$, such that, $\hat{g}_1(x;v\setminus\hat{v}) \neq 0$, we have:

$$\hat{v}_1 = -\frac{\sum_{i=2}^{k} \hat{v}_i \cdot \hat{g}_i(x;v\setminus\hat{v})}{\hat{g}_1(x;v\setminus\hat{v})} \tag{3}$$

where k is the dimension of $\hat{g}(x;v\setminus\hat{v})$. We notice that since the left hand side of Eq. 3 is a continuous function, the set $\{v \mid \hat{g}_1(x;v\setminus\hat{v}) \neq 0, g(x;v)=0\}$ can be represented as a graph of a continuous function, where $\hat{v}_1$ satisfies Eq. 3. Therefore, it is of measure zero. Hence, $\{w \mid g(x;w)=0\}$ is of measure zero as well. □

Lemma 2. *Let $f(x;w)$ be a neural network (e.g., vanilla ReLU network, ResNet or DenseNet). Let x be a non-zero vector. Then, the set $\left\{w \mid J^{\mathbf{k}} = \frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\}$ is of measure 1.*

Proof. It holds that:

$$J^{\mathbf{k}} = \frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}} + \frac{\partial f^{c}_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}} \tag{4}$$

We would like to prove that the set of w, such that, $\frac{\partial f^c_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}} = 1$ is of measure 1.

First, we consider that the set of weights $w_{\gamma,l}$ within the expression $f^c_{\mathbf{k}}(x;w) = \sum_{\gamma \in S \setminus S_{\mathbf{k}}} c_\gamma z_\gamma \prod_{l=1}^{|\gamma|} w_{\gamma,l}$ is disjoint to the set of weights $w^k_{i,j}$ in $W^{\mathbf{k}}$, since the complement $f^c_{\mathbf{k}}(x;w)$ sums over the paths γ that skip $W^{\mathbf{k}}$. We note that z_γ is a binary function that indicates whether the neurons along the path γ are activated or not. Therefore, for any $\gamma \in S \setminus S_{\mathbf{k}}$, we have: $\frac{\partial z_\gamma}{\partial W^{\mathbf{k}}} = 0$ for every w, such that, the pre-activations of each neuron along the path γ are non-zero (otherwise, the gradient is undefined). By Lem. 1, the complement of this set (i.e., all w, such that, the pre-activation of at least one neuron along the path γ is zero) is of measure zero. Therefore, we conclude that $\frac{\partial z_\gamma}{\partial W^{\mathbf{k}}} = 0$ holds almost surely. Since this is true for all $\gamma \in S \setminus S_{\mathbf{k}}$, we conclude that $\frac{\partial f^c_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}} = 0$ almost surely. □

Lemma 3. *Let $f(x;w)$ be a neural network (e.g., vanilla ReLU network, ResNet or DenseNet). Let x be a non-zero vector. Then,*

$$\mathbb{E}[\|J^{\mathbf{k}}\|_2^p] = \mathbb{E}\left[\left\|\frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\|_2^p\right] \tag{5}$$

Proof. By Lem. 2, the set $\left\{w \mid J^{\mathbf{k}} = \frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\}$ is of measure 1. Therefore, since w is distributed according to a continuous distribution, we have the desired equation: $\mathbb{E}[\|J_k\|_2^p] = \mathbb{E}\left[\|\partial f_{\mathbf{k}}(x;w)/\partial W^{\mathbf{k}}\|_2^p\right]$. □

4 PROOFS OF THE MAIN RESULTS

We make use of the following propositions and definitions to aid in the proofs of Thms. 1 and 2.

Proposition 1. *Given a random vector $w = [w_1 ... w_n]$ such that each component is identically and symmetrically distributed i.i.d random variable with moments $\mathbb{E}[w_1^m] = c_m$ (e.g., $c_0 = 1, c_1 = 0$), a set of non negative integers $m_1, ..., m_l$, such that, $\sum_{i=1}^l m_i$ is even, and a random binary variable $z \in \{0,1\}$, such that, $p(z \mid w) = 1 - p(z \mid -w)$, then it holds that:*

$$\mathbb{E}\left[\prod_{i=1}^l w_i^{m_i} z\right] = \frac{\prod_{i=1}^l c_{m_i}}{2} \tag{6}$$

Proof. We have:

$$\begin{aligned}\prod_{i=1}^l c_{m_i} &= \int_w \prod_{i=1}^l w_i^{m_i} p(w)\,\mathrm{d}w \\ &= \int_{w|z=1} \prod_{i=1}^l w_i^{m_i} p(w)\,\mathrm{d}w + \int_{w|z=0} \prod_{i=1}^l w_i^{m_i} p(w)\,\mathrm{d}w \\ &= \int_{w|z=1} \prod_{i=1}^l w_i^{m_i} p(w)\,\mathrm{d}w + \int_{w|z=1} \prod_{i=1}^l (-w_i)^{m_i} p(w)\,\mathrm{d}w \\ &= \int_w \prod_{i=1}^l w_i^{m_i} z \cdot p(w)\,\mathrm{d}w + \int_w \prod_{i=1}^l (-w_i)^{m_i} z \cdot p(w)\,\mathrm{d}w\end{aligned} \tag{7}$$

Since $\sum_{i=1}^l m_i$ is even, it follows that:

$$\int_w \prod_{i=1}^l (-w_i)^{m_i} z \cdot p(w)\,\mathrm{d}w = \int_w \prod_{i=1}^l w_i^{m_i} z \cdot p(w)\,\mathrm{d}w \tag{8}$$

Therefore,

$$\prod_{i=1}^l c_{m_i} = 2 \int_w \prod_{i=1}^l w_i^{m_i} z \cdot p(w)\,\mathrm{d}w \tag{9}$$

Put differently,

$$\frac{\prod_{i=1}^l c_{m_i}}{2} = \int_w \prod_{i=1}^l w_i^{m_i} z \cdot p(w)\,\mathrm{d}w = \mathbb{E}\left[\prod_{i=1}^l w_i^{m_i} z\right] \tag{10}$$

□

Proposition 2. *Given a random vector $w = [w_1, ..., w_n]$, such that,its components are i.i.d symmetrically distributed random variable with moments $\mathbb{E}[w_i^m] = c_m$ ($c_0 = 1, c_1 = 0$), two sets of non negative integers $m_1, ..., m_l$, $n_1, ..., n_l$, such that, $\sum_{i=1}^l m_i$, $\sum_{i=1}^l n_i$ are even, $\forall\ i \in [l] : m_i \geq n_i$, and a random binary variable $z \in \{0, 1\}$, such that $p(z \mid w) = 1 - p(z \mid -w)$, then it holds that:*

$$\mathbb{E}\left[\frac{1}{w_i^{n_i}}\prod_{i=1}^{l} w_i^{m_i} z\right] = \frac{\prod_{i=1}^{l} c_{m_i - n_i}}{2} \tag{11}$$

Proof. Follows immediately from Prop. 1 since $\sum_i (m_i - n_i)$ is even. □

Definition 1 (ResNet path parametrization). *Let $f(x; w)$ be a ResNet with two layer residual branches ($m = 2$). A path from input to output γ in f, defines a product of weights along the path denoted by:*

$$P_\gamma = \prod_{l=0}^{L+1} p_{\gamma,l} \tag{12}$$

where:

$$p_{\gamma,l} = \begin{cases} 1 & l \notin \gamma \\ w^1_{\gamma,l} z_{\gamma,l} w^2_{\gamma,l} & l \in \gamma, 0 < l \leq L \\ w_{\gamma,l} & l = \{0, L+1\} \end{cases} \tag{13}$$

Here, $w^1_{\gamma,l}, w^2_{\gamma,l}$ are weights associated with residual branch l, $w_{\gamma,0}, w_{\gamma,L+1}$ belong to the first and last linear projection matrices W^0, W^{L+1}, and $z_{\gamma,l}$ is the binary activation variable relevant for weight $w^1_{\gamma,l}$. (Note that $z_{\gamma,l}$ depends on $w^1_{\gamma,l}$, but not on $w^2_{\gamma,l}$). $l \notin \gamma$ indicates if layer l is skipped.

Definition 2 (DenseNet path parametrization). *Let $f(x; w)$ be a DenseNet. A path γ from input in to output in f, defines a product of weights along the path denoted by:*

$$P_\gamma = \prod_{l=0}^{L+1} p_{\gamma,l} \tag{14}$$

where:

$$p_{\gamma,l} = \begin{cases} 1 & l \notin \gamma \\ w_{\gamma,l} z_{\gamma,l} & l \in \gamma, 0 < l \leq L \\ w_{\gamma,l} & l = \{0, L+1\} \end{cases} \tag{15}$$

Here, $w_{\gamma,l}$ is a weight associated with layer l, $w_{\gamma,0}, w_{\gamma,L+1}$ belong to the first and last linear projection matrices W^0, W^{L+1}, and $z_{\gamma,l}$ is the binary activation variable relevant for weight $w_{\gamma,l}$. The notation $l \notin \gamma$ indicates that the layer l is skipped.

Similarly, we denote $z^{(\mathbf{k})}_{\gamma,l}$, $p^{(\mathbf{k})}_{\gamma,l}$ and $P^{(\mathbf{k})}_\gamma$ to be the same quantities as $z_{\gamma,l}$, $p_{\gamma,l}$ and P_γ for the network $f_{(\mathbf{k})}$ instead of f.

Proposition 3. *Let $f(x; w)$ be a ResNet/DenseNet/ANN. For any set of even m paths from input to output $\{\gamma^i\}_{i=1}^m$, it holds that:*

$$\mathbb{E}\left[\prod_{i=1}^{m} P_{\gamma^i}\right] = \begin{cases} \prod_{l=0}^{L+1}\left(\mathbb{E}\left[\prod_{i=1}^m p_{\gamma^i,l} \mid \sum_{h=0}^{l-1} \|q^h\|_2 > 0\right]\right) & f(x; w) \text{ is DenseNet} \\ \prod_{l=0}^{L+1}\left(\mathbb{E}\left[\prod_{i=1}^m p_{\gamma^i,l} \mid \|y^{l-1}\|_2 > 0\right]\right) & f(x; w) \text{ is ResNet or ANN} \end{cases} \tag{16}$$

Proof. We prove the claim for DenseNets. The extension to ANNs and ResNets is trivial, and requires no further arguments. We have that:

$$\mathbb{E}\left[\prod_{i=1}^{m} P_{\gamma^i}\right] = \mathbb{E}\left[\prod_{l=0}^{L+1}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right] \tag{17}$$

From the linearity of the last layer, it follows that:

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{m} P_{\gamma^i}\right] &= \mathbb{E}\left[\prod_{l=0}^{L}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L+1}\right]\\ &= \mathbb{E}\left[\prod_{l=0}^{L}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\mathbb{E}\left[\prod_{i=1}^{m} w_{\gamma^i,L+1}\right]\end{aligned}\tag{18}$$

We denote by $\{w_u^{L+1}\}_{u=1}^{s}$ the set of $s \leq m$ unique weights in $\{w_{\gamma^i,L+1}\}_{i=1}^{m}$, with corresponding multiplicities $\{m_u^{L+1}\}_{u=1}^{s}$, such that, $\sum_{u=1}^{s} m_u^{L+1} = m$. It follows that:

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{m} P_{\gamma^i}\right] &= \mathbb{E}\left[\prod_{l=0}^{L}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\mathbb{E}\left[\prod_{u}\left(w_u^{L+1}\right)^{m_u^{L+1}}\right]\\ &= \mathbb{E}\left[\prod_{l=0}^{L}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\prod_{u} c_{m_u^{L+1}}\end{aligned}\tag{19}$$

where $c_{m_u^{L+1}}$ is the m_u^{L+1}'th moment of a normal distribution.

Since the computations done by all considered architectures form a Markov chain, such that, the output of any layer depends only on the set R^{l-1} of weights in the previous layers, we have that:

$$\mathbb{E}\left[\prod_{l=0}^{L}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right] = \mathbb{E}\left[\prod_{l=0}^{L-1}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle| R^{L-1}\right]\right]\tag{20}$$

And also,

$$\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle| R^{L-1}\right] = \mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle| R^{L-1}\right] = \mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle| q^0,...,q^{L-1}\right]\tag{21}$$

We note that the pre-activations y^L conditioned on $q^0,...,q^{L-1}$ are distributed according to zero mean i.i.d Gaussian variables. In addition, the coordinates of $q^L = 2\phi(y^L)$ are i.i.d distributed. We denote by $\{z_u\}_{u=1}^{s}$ the set of unique activation variables in the set $\{z_{\gamma^i,L}\}_{i=1}^{m}$. For each z_u, we denote by $\{w_{u,v}^L\}$ the set of unique weights in $\{w_{\gamma^i,L}\}$ multiplying z_u, with corresponding multiplicities $m_{u,v}^L$, such that, $\sum_{u,v} m_{u,v}^L = m$, and $\sum_v m_{u,v}^L = m_u^{L+1}$. Note that, from the symmetry of the normal distribution, it holds that odd moments vanish, and so we only need to consider even m_u^{L+1} for all u. From the independence of the set $\{z_u\}$, the expectation takes a factorized form:

$$\begin{aligned}\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle| q^0...q^{L-1}\right] &= \mathbb{1}\left[\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\cdot\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L} \mid q^0,...,q^{L-1}\right]\\ &= \mathbb{1}\left[\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\cdot\prod_{u=1}^{s}\mathbb{E}\left[z_u\prod_{v}(w_{u,v}^L)^{m_{u,v}^L}\middle| q^0,...,q^{L-1}\right]\end{aligned}\tag{22}$$

Using Prop. 1:

$$\begin{aligned}&\prod_{u=1}^{s}\mathbb{E}\left[z_u\prod_{v}(w_{u,v}^L)^{m_{u,v}^L}\middle| q^0...q^{L-1}\right]\\ =&\mathbb{1}\left[\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\cdot\prod_{u=1}^{s}\left(\frac{\prod_v c_{m_{u,v}^L}}{2}\right)\\ =&\mathbb{1}\left[\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\cdot\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle|\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\end{aligned}\tag{23}$$

It then follows:

$$\begin{aligned}\mathbb{E}\left[\prod_{l=0}^{L}(\prod_{i=1}^{m} p_{\gamma^i,l})\right] &= \mathbb{E}\left[\mathbb{1}\left[\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\cdot\prod_{l=0}^{L-1}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\prod_{u=1}^{s}\left(\frac{\prod_v c_{m_{u,v}^L}}{2}\right)\\ &= \mathbb{E}\left[\prod_{l=0}^{L-1}\left(\prod_{i=1}^{m} p_{\gamma^i,l}\right)\right]\cdot\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,L}\middle|\sum_{l=0}^{L-1}\|q^l\|_2 > 0\right]\end{aligned}\tag{24}$$

Recursively applying the above completes the proof. □

Theorem 1. *Let $f(x;w)$ be a ResNet/DenseNet. Then, for any non-negative even integer m, we have:*

$$\forall\, \mathbf{k}: \; \mathbb{E}_w\left[(f_{(\mathbf{k})}(x;w))^m\right] = \mathbb{E}_w\left[(f_{\mathbf{k}}(x;w))^m\right] \tag{25}$$

Proof. We present the proof using the DenseNet path parameterization. Extending to ResNet parameterization is trivial and requires no additional arguments. We aim to show that for any even integer $m > 0$, and $\forall\, \mathbf{k} = \{l_k, h_k\}$:

$$\mathbb{E}\left[(f_{(\mathbf{k})}(x;w))^m\right] = \mathbb{E}\left[(f_{\mathbf{k}}(x;w))^m\right] \tag{26}$$

The output $f_{\mathbf{k}}(x;w)$ can be expressed in the following manner:

$$f_{\mathbf{k}}(x;w) = \sum_{\gamma \in S_{\mathbf{k}}} c_\gamma \prod_{l=0}^{L+1} p_{\gamma,l} \tag{27}$$

Since the output $f_{(\mathbf{k})}(x;w)$ is composed of products of weights and activations along the same paths $\gamma \in S_{\mathbf{k}}$ as $f_{\mathbf{k}}$ (with different activation variables), we only need to prove the following: for any weight matrix $W^{\mathbf{k}}$, and a set of m paths $\gamma^1, ..., \gamma^m \in S_{\mathbf{k}}$, it holds that:

$$\mathbb{E}\left[\prod_{i=1}^{m} P_{\gamma^i}\right] = \mathbb{E}\left[\prod_{i=1}^{m} P^{(\mathbf{k})}_{\gamma^i}\right] \tag{28}$$

Using Prop. 3:

$$\prod_{l=0}^{L+1}\left(\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,l} \,\middle|\, \sum_{h=1}^{l-1} \|q^h\|_2 > 0\right]\right) = \prod_{l=0}^{L+1}\left(\mathbb{E}\left[\prod_{i=1}^{m} p^{\mathbf{k}}_{\gamma^i,l} \,\middle|\, \sum_{h=1}^{l-1} \|q^h_{(\mathbf{k})}\|_2 > 0\right]\right) \tag{29}$$

Note that for both the full and reduced architectures, flipping the sign of all the weights in layer l will flip the ensuing activation variables (except for a set of measure zero defined by $\sum_{l=0}^{l_k-1} W^{l_k,l} q^l = 0$, which does not affect the expectation. And so, using Prop. 1 along with Eq. 23:

$$\mathbb{E}\left[\prod_{i=1}^{m} p_{\gamma^i,l} \,\middle|\, \sum_{h=1}^{l-1} \|q^h\|_2 > 0\right] = \mathbb{E}\left[\prod_{i=1}^{m} p^{\mathbf{k}}_{\gamma^i,l} \,\middle|\, \sum_{h=1}^{l-1} \|q^h_{(\mathbf{k})}\|_2 > 0\right] \tag{30}$$

completing the proof. □

Theorem 2. *Let $f(x;w)$ be a ResNet/DenseNet. Then, we have for all k:*

1. $\mathbb{E}_w\left[\|J^{\mathbf{k}}\|_2^2\right] = \mathbb{E}_w\left[(f_{(\mathbf{k})}(x;w))^2\right]$.
2. $\frac{\mathbb{E}_w\left[(f_{(\mathbf{k})}(x;w))^4\right]}{3} \leq \mathbb{E}_w\left[\|J^{\mathbf{k}}\|_2^4\right] \leq \mathbb{E}_w\left[(f_{(\mathbf{k})}(x;w))^4\right]$.

Proof. We present the proof using the DenseNet path parameterization. Extending to ResNet parameterization is trivial and requires no additional arguments. Neglecting scaling coefficients for notational simplicity, let $\mathbf{k} = (l_k, h_k)$ be an index of a weight matrix $W^{\mathbf{k}}$ in $f(x;w)$, by Lem. 3, we have:

$$\mathbb{E}\left[\|J^{\mathbf{k}}\|_2^2\right] = \mathbb{E}\left[\left\|\frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\|_2^2\right] = \sum_{i,j}\mathbb{E}\left[\left(\sum_{\gamma \in S \text{ s.t: } w^{\mathbf{k}}_{i,j} \in \gamma} \frac{1}{w^{\mathbf{k}}_{i,j}} P_\gamma\right)^2\right] \tag{31}$$

where γ s.t: $w^{\mathbf{k}}_{i,j} \in \gamma$ denotes a path that includes the weight $w^{\mathbf{k}}_{i,j}$. From Prop. 3, the expectation is factorized as follows:

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\|_2^2\right]\\ =&\sum_{i,j}\sum_{\gamma \in S \text{ s.t: } w^{\mathbf{k}}_{i,j} \in \gamma} \mathbb{E}\left[\left(\frac{1}{w^{\mathbf{k}}_{i,j}} p_{\gamma,l_k}\right)^2 \,\middle|\, \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] \cdot \prod_{l \neq l_k} \mathbb{E}\left[(p_{\gamma,l_k})^2 \,\middle|\, \sum_{h=0}^{l-1} \|q^h\|_2 > 0\right]\end{aligned} \tag{32}$$

Using Props. 1 and 2, for all $\gamma \in S$, such that, $w^{\mathbf{k}}_{i,j} \in \gamma$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{1}{w^{\mathbf{k}}_{i,j}} p_{\gamma,l_k}\right)^2 \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] \\
=&\mathbb{E}\left[\left(\frac{w^{\mathbf{k}}_{i,j} z_{\gamma,l_k}}{w^{\mathbf{k}}_{i,j}}\right)^2 \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] = 1/2 = \mathbb{E}\left[(p_{\gamma,l_k})^2 \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right]
\end{aligned} \tag{33}$$

Inserting into Eq. 32, and using Thm. 1 proves the first claim.

Next we would like to prove the second claim. By Lem. 1, we have:

$$\begin{aligned}
\mathbb{E}\left[\|J^{\mathbf{k}}\|_2^4\right] =&\mathbb{E}\left[\left\|\frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\|_2^2 \cdot \left\|\frac{\partial f_{\mathbf{k}}(x;w)}{\partial W^{\mathbf{k}}}\right\|_2^2\right] \\
=&\sum_{i,j}\sum_{i',j'} \mathbb{E}\left[\left(\sum_{\gamma \text{ s.t } w^{\mathbf{k}}_{i,j} \in \gamma} \frac{1}{w^{\mathbf{k}}_{i,j}} P_\gamma\right)^2 \left(\sum_{\gamma \text{ s.t } w^{\mathbf{k}}_{i',j'} \in \gamma} \frac{1}{w^{\mathbf{k}}_{i',j'}} P_\gamma\right)^2\right] \\
=&\sum_{i,i',j,j'} \mathbb{E}\left[\frac{1}{(w^{\mathbf{k}}_{i,j})^2 (w^{\mathbf{k}}_{i',j'})^2} \sum_{\gamma^1,\gamma^2 \text{ s.t } w^{\mathbf{k}}_{i,j} \in \gamma^1,\gamma^2} \quad \sum_{\gamma^3,\gamma^4 \text{ s.t } w^{\mathbf{k}}_{i',j'} \in \gamma^3,\gamma^4} P_{\gamma^1} P_{\gamma^2} P_{\gamma^3} P_{\gamma^4}\right]
\end{aligned} \tag{34}$$

By applying Prop. 3, the expectation is factorized as follows:

$$\begin{aligned}
&\mathbb{E}\left[\|J^{\mathbf{k}}\|_2^4\right] \\
=& \sum_{\substack{i,i',j,j' \\ \gamma^1,\gamma^2 \text{ s.t } w^{\mathbf{k}}_{i,j} \in \gamma^1,\gamma^2 \\ \gamma^3,\gamma^4 \text{ s.t } w^{\mathbf{k}}_{i',j'} \in \gamma^3,\gamma^4}} \left[\mathbb{E}\left[\frac{\prod_{h=1}^4 p_{\gamma^h,l_k}}{(w^{\mathbf{k}}_{i,j})^2 (w^{\mathbf{k}}_{i',j'})^2} \middle| \sum_{h=0}^{l_k-1} \|q^h\| > 0\right] \cdot \prod_{l \neq l_k} \mathbb{E}\left[\prod_{h=1}^4 p_{\gamma^h,l} \middle| \sum_{h=0}^{l-1} \|q^h\| > 0\right]\right]
\end{aligned} \tag{35}$$

Using Props. 1 and 2, for all γ^1, γ^2, such that, $w^k_{i,j} \in \gamma^1$ and $w^{\mathbf{k}}_{i',j'} \in \gamma^2$, we have:

$$\begin{aligned}
&\mathbb{E}\left[\frac{\prod_{h=1}^4 p_{\gamma^h,k}}{(w^{\mathbf{k}}_{i,j})^2 (w^{\mathbf{k}}_{i',j'})^2} \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] \\
=&\mathbb{E}\left[\frac{(w^{\mathbf{k}}_{i,j})^2 (w^{\mathbf{k}}_{i',j'})^2 z_{\gamma^1,k} z_{\gamma^2,k}}{(w^{\mathbf{k}}_{i,j})^2 (w^{\mathbf{k}}_{i',j'})^2} \middle| \sum_{h=0}^{k-1} \|q^h\|_2 > 0\right] \\
=&\begin{cases} 1/2 & w^{\mathbf{k}}_{i,j} \equiv w^{\mathbf{k}}_{i',j'} \\ 1/4 & otherwise \end{cases} \\
=&\begin{cases} \frac{1}{3}\mathbb{E}\left[\prod_{h=1}^4 p_{\gamma^h,k} \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] & w^{\mathbf{k}}_{i,j} \equiv w^{\mathbf{k}}_{i',j'} \\ \mathbb{E}\left[\prod_{h=1}^4 p_{\gamma^h,k} \middle| \sum_{h=0}^{l_k-1} \|q^h\|_2 > 0\right] & otherwise \end{cases}
\end{aligned} \tag{36}$$

Inserting into Eq. 35 proves the second claim. □

We use the following proposition to aid in the proofs of Thms. 3 and 4.

Proposition 4. *Let $f(x;w)$ be a vanilla fully connected ReLU network, with intermediate outputs given by:*

$$\forall\, 0 \leq l \leq L: \; y^l = \sqrt{2}\phi\left(\frac{1}{\sqrt{n_{l-1}}} W^l y^{l-1}\right) \tag{37}$$

where the weight matrices $W^l \in \mathbb{R}^{n_l \times n_{l-1}}$ are normally distributed. Then, the following holds at initialization:

$$\begin{aligned}\mathbb{E}\left[\|y^l\|_2^2\right] &= \frac{n_l}{n_{l-1}}\mathbb{E}\left[\|y^{l-1}\|_2^2\right]\\ \mathbb{E}\left[\|y^l\|_2^4\right] &= \frac{n_l(n_l+5)}{n_{l-1}^2}\mathbb{E}\left[\|y^{l-1}\|_2^4\right]\end{aligned} \tag{38}$$

Proof. Absorbing the scale $\sqrt{\frac{2}{n_{l-1}}}$ into the weights, we denote by Z^l the diagonal matrix holding in its diagonal the activation variables z_j^l for unit j in layer l, and so we have:

$$y^l = Z^l W^l y^{l-1} \tag{39}$$

Conditioning on $R^{l-1} = \{W^1, ..., W^{l-1}\}$ and taking expectation:

$$\begin{aligned}\mathbb{E}\left[\|y^l\|_2^2 \mid R^{l-1}\right] &= {y^{l-1}}^\top \mathbb{E}\left[{W^l}^\top Z^l W^l\right] y^{l-1}\\ &= \sum_{j=1}^{n_l}\sum_{i_1,i_2=1}^{n_{l-1}} y_{i_1}^{l-1} y_{i_2}^{l-1}\mathbb{E}\left[w_{i_1,j}^l w_{i_2,j}^l z_j^l \mid R^{l-1}\right]\end{aligned} \tag{40}$$

From Prop. 1, it follows that:

$$\mathbb{E}\left[\|y^l\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|y^l\|_2^2 \mid R^{l-1}\right]\right] = \frac{n_L}{n_{L-1}}\mathbb{E}\left[\|y^{l-1}\|_2^2\right] \tag{41}$$

Similarly:

$$\begin{aligned}\mathbb{E}\left[\|y^l\|_2^4 \mid R^{l-1}\right] &= \mathbb{E}\left[\left({y^{l-1}}^\top {W^l}^\top Z^l W^l y^{l-1}\right)^2 \middle| R^{l-1}\right]\\ &= \sum_{j_1,j_2,i_1,i_2,i_3,i_4}\prod_{t=1}^4 y_{i_t}^{l-1} \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l z_{j_1}^l z_{j_2}^l \mid R^{l-1}\right]\end{aligned} \tag{42}$$

From Prop. 1, and the independence of the activation variables conditioned on R^{l-1}:

$$\begin{aligned}&\sum_{j_1,j_2,i_1,i_2,i_3,i_4}\prod_{t=1}^4 y_{i_t}^{l-1} \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l z_{j_1}^l z_{j_2}^l | R^{l-1}\right]\\ =&\sum_{j_1,j_2,i_1,i_2,i_3,i_4}\prod_{t=1}^4 y_{i_t}^{l-1} \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l z_{j_1}^l z_{j_2}^l | R^{l-1}\right]\\ &\cdot \Big(\mathbb{1}_{j_1=j_2,i_1=i_2=i_3=i_4} + \mathbb{1}_{j_1=j_2,i_1=i_2,i_3=i_4,i_1\neq i_3}\\ &\quad + \mathbb{1}_{j_1=j_2,i_1=i_3,i_2=i_1,i_2\neq i_3} + \mathbb{1}_{j_1=j_2,i_1=i_4,i_2=i_3,i_1\neq i_2} + \mathbb{1}_{j_1\neq j_2,i_1=i_2,i_3=i_4}\Big)\end{aligned} \tag{43}$$

and so:

$$\begin{aligned}\mathbb{E}\left[\|y^l\|_2^4\right] =& \frac{6n_l}{n_{l-1}^2}\sum_i \mathbb{E}\left[(y_i^{l-1})^4\right] + \frac{6n_l}{n_{l-1}^2}\sum_{i_1\neq i_2}\mathbb{E}\left[(y_{i_1}^{l-1})^2(y_{i_2}^{l-1})^2\right]\\ &+ \frac{n_l(n_l-1)}{n_{l-1}^2}\sum_{i_1,i_2}\mathbb{E}\left[(y_{i_1}^{l-1})^2(y_{i_2}^{l-1})^2\right]\\ =& \frac{n_l(n_l+5)}{n_{l-1}^2}\mathbb{E}\left[\|y^{l-1}\|_2^4\right]\end{aligned} \tag{44}$$

proving the claim.

□

Proposition 5. *For a vanilla fully connected linear network, with intermediate outputs given by:*

$$\forall\, 0 \leq l \leq L:\ y^l = \frac{1}{\sqrt{n_{l-1}}} W^l y^{l-1} \tag{45}$$

where the weight matrices $W^l \in \mathbb{R}^{n_l \times n_{l-1}}$ are normally distributed, the following holds at initialization:

$$\begin{aligned}
\mathbb{E}\left[\|y^l\|_2^2\right] &= \frac{n_l}{n_{l-1}} \mathbb{E}\left[\|y^{l-1}\|_2^2\right] \\
\mathbb{E}\left[\|y^l\|_2^4\right] &= \frac{n_l(n_l+2)}{n_{l-1}^2} \mathbb{E}\left[\|y^{l-1}\|_2^4\right]
\end{aligned} \tag{46}$$

Proof. The proof follows immediately from the derivation of Prop. 4, and will be omitted for brevity.

Proposition 6. *Let $f(x;w)$ be a vanilla fully connected linear network, with intermediate outputs given by:*

$$\forall\, 0 \leq l \leq L:\ y^l(x) = \frac{1}{\sqrt{n_{l-1}}} W^l y^{l-1}(x) \tag{47}$$

where the weight matrices $W^l \in \mathbb{R}^{n_l \times n_{l-1}}$ are normally distributed. Then, the following holds at initialization:

$$\begin{aligned}
\mathbb{E}\left[(y^l(x'))^\top \cdot y^l(x)\right] &= \frac{n_l}{n_{l-1}} \mathbb{E}\left[(y^{l-1}(x'))^\top \cdot y^{l-1}(x)\right] \\
\mathbb{E}\left[(y^l(x')^\top y^l(x))^2\right] &= \frac{1}{n} \mathbb{E}\left[((y^{l-1}(x'))^\top y^{l-1}(x))^2\right] + \mathbb{E}\left[\|y^{l-1}(x')\|^2 \|y^{l-1}(x)\|^2\right] \\
\mathbb{E}\left[\|y^l(x')\|^2 \|y^l(x)\|^2\right] &= \mathbb{E}\left[((y^{l-1}(x'))^\top y^{l-1}(x))^2\right] + \frac{1}{n} \mathbb{E}\left[\|y^{l-1}(x')\|^2 \|y^{l-1}(x)\|^2\right].
\end{aligned} \tag{48}$$

Proof. Absorbing the scale $\sqrt{\frac{1}{n_{l-1}}}$ into the weights, we can write instead:

$$y^l(x) = W^l y^{l-1}(x) \tag{49}$$

Conditioning on $R^{l-1} = \{W^1, ..., W^{l-1}\}$ and taking expectation:

$$\begin{aligned}
\mathbb{E}\left[(y^l(x'))^\top \cdot y^l(x) \mid R^{l-1}\right] &= y^{l-1}(x')^\top \mathbb{E}\left[{W^l}^\top W^l\right] y^{l-1}(x) \\
&= \sum_{j=1}^{n_l} \sum_{i_1, i_2 = 1}^{n_{l-1}} y_{i_1}^{l-1}(x') y_{i_2}^{l-1}(x) \mathbb{E}\left[w_{i_1,j}^l w_{i_2,j}^l \mid R^{l-1}\right]
\end{aligned} \tag{50}$$

From Prop. 1, it follows that:

$$\mathbb{E}\left[(y^l(x'))^\top \cdot y^l(x)\right] = \mathbb{E}\left[\mathbb{E}\left[(y^l(x'))^\top \cdot y^l(x) \mid R^{l-1}\right]\right] = \frac{n_L}{n_{L-1}} \mathbb{E}\left[y^{l-1}(x')^\top \cdot y^{l-1}(x)\right] \tag{51}$$

Similarly:

$$\begin{aligned}
\mathbb{E}\left[((y^l(x'))^\top \cdot y^l(x))^2 \mid R^{l-1}\right] &= \mathbb{E}\left[\left(y^{l-1}(x')^\top {W^l}^\top W^l y^{l-1}(x)\right)^2 \middle| R^{l-1}\right] \\
&= \sum_{j_1, j_2, i_1, i_2, i_3, i_4} y_{i_1}^{l-1}(x') \cdot y_{i_2}^{l-1}(x') \cdot y_{i_3}^{l-1}(x) \cdot y_{i_4}^{l-1}(x) \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l \mid R^{l-1}\right]
\end{aligned} \tag{52}$$

We denote: $Q_{i_1,...,i_4}^l := y_{i_1}^l(x') \cdot y_{i_2}^l(x') \cdot y_{i_3}^l(x) \cdot y_{i_4}^l(x)$. From Prop. 1, and the independence of the activation variables conditioned on R^{l-1}:

$$\begin{aligned}
&\sum_{j_1, j_2, i_1, i_2, i_3, i_4} Q_{i_1,...,i_4}^l \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l | R^{l-1}\right] \\
= &\sum_{j_1, j_2, i_1, i_2, i_3, i_4} Q_{i_1,...,i_4}^l \cdot \mathbb{E}\left[w_{i_1,j_1}^l w_{i_2,j_1}^l w_{i_3,j_2}^l w_{i_4,j_2}^l | R^{l-1}\right] \\
&\cdot \Big(\mathbb{1}_{j_1=j_2, i_1=i_2=i_3=i_4} + \mathbb{1}_{j_1=j_2, i_1=i_2, i_3=i_4, i_1 \neq i_3} \\
&\quad + \mathbb{1}_{j_1=j_2, i_1=i_3, i_2=i_1, i_2 \neq i_3} + \mathbb{1}_{j_1=j_2, i_1=i_4, i_2=i_3, i_1 \neq i_2} + \mathbb{1}_{j_1 \neq j_2, i_1=i_2, i_3=i_4}\Big)
\end{aligned} \tag{53}$$

After some quick calculations we have:

$$\mathbb{E}\left[(y^l(x')^\top y^l(x))^2\right] = \frac{1}{n}\mathbb{E}\left[((y^{l-1}(x'))^\top y^{l-1}(x))^2\right] + \mathbb{E}\left[\|y^{l-1}(x')\|^2\|y^{l-1}(x)\|^2\right]. \tag{54}$$

proving the claim.

□

Theorem 3. *Let $f(x; w)$ be a depth L, constant width ResNet with residual branches of depth m (with $n_0', n_l, n_{l,h} = n$ for all $l \in [L]$ and $h \in [m]$), with positive initialization constants $\{\alpha_l\}_{l=1}^L$. Then, there exists a constant $C > 0$ such that:*

$$\max\left[1, \frac{\sum_{\mathbf{u}} \alpha_{l_u}^2}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}} \cdot \xi\right] \leq \eta(n, L) \leq \xi \tag{55}$$

where:

$$\xi = \exp\left[\frac{5m}{n} + \frac{C}{n}\sum_{l=1}^{L}\frac{\alpha_l}{1+\alpha_l}\right] \cdot (1 + \mathcal{O}(1/n)) \tag{56}$$

Proof. Using the result of Thm. 2, and using Cauchy–Schwartz inequality, an upper bound to η can be derived:

$$\begin{aligned}
\eta &= \frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathring{\mathcal{K}}_L^R(x,x)^2} \\
&= \frac{\sum_{\mathbf{u},\mathbf{v}} \mathbb{E}[\|J^{\mathbf{u}}\|_2^2 \cdot \|J^{\mathbf{v}}\|_2^2]}{\mathring{\mathcal{K}}_L^R(x,x)^2} \\
&\leq \frac{\sum_{\mathbf{u},\mathbf{v}} \sqrt{\mathbb{E}[\|J^{\mathbf{u}}\|_2^4] \cdot \mathbb{E}[\|J^{\mathbf{v}}\|_2^2]}}{\mathring{\mathcal{K}}_L^R(x,x)^2} \\
&\leq \frac{\sum_{\mathbf{u},\mathbf{v}} \sqrt{\mathbb{E}[\|f_{(\mathbf{u})}(x;w)\|_2^4] \cdot \mathbb{E}[\|f_{(\mathbf{v})}(x;w)\|_2^2]}}{\mathring{\mathcal{K}}_L^R(x,x)^2}
\end{aligned} \tag{57}$$

The lower bound is similarly derived using Thm. 2:

$$\eta \geq \frac{\sum_{\mathbf{k}} \mathbb{E}[\|J^{\mathbf{k}}\|_2^4]}{\mathring{\mathcal{K}}_L^R(x,x)^2} \geq \frac{1}{3} \cdot \frac{\sum_{\mathbf{k}} \mathbb{E}[\|f_{(\mathbf{k})}(x;w)\|_2^4]}{\mathring{\mathcal{K}}_L^R(x,x)^2} \tag{58}$$

The asymptotic behaviour of η is therefore governed by the propagation of the fourth moment $\mathbb{E}[\|y_{(\mathbf{k})}^l\|_2^4]$ through the model.

In the following proof, for the sake of notation simplicity, we omit the notation $\mathbf{k} = (l_k, h_k)$ in $y_{(\mathbf{k})}^l$, and assume that y^l stands for the reduced network $y_{(\mathbf{k})}^l$. The recursive formula for the intermediate outputs of the reduced network are given by:

$$y^l = \begin{cases} y^{l-1} + \sqrt{\alpha_l}y^{l-1,m} & 0 < l \leq L, l \neq l_k \\ \sqrt{\alpha_l}y^{l-1,m} & l = l_k \end{cases} \tag{59}$$

where:

$$y^{l-1,h} = \begin{cases} \sqrt{\frac{1}{n}}W^{l,h}q^{l-1,h-1} & 1 < h \leq m \\ \sqrt{\frac{1}{n}}W^{l,h}y^{l-1} & h = 1 \end{cases} \tag{60}$$

with $q^{l-1,h} = \sqrt{2}\phi(y^{l-1,h})$.

Using the results of Props. 4 and 5, for layer L, we have:

$$\begin{aligned}
\mathbb{E}\left[\|y^L\|_2^2\right] =& \mathbb{E}\left[\|y^{L-1}\|_2^2\right] + \frac{\alpha_L}{n}\mathbb{E}\left[{y^{L-1,m-1}}^\top {W^{L,m}}^\top W^{L,m} y^{L-1,m-1}\right] \\
=& \mathbb{E}\left[\|y^{L-1}\|_2^2\right] + \alpha_L \mathbb{E}\left[\|y^{L-1,m-1}\|_2^2\right] \\
=& \mathbb{E}\left[\|y^{L-1}\|_2^2\right] \cdot (1+\alpha_L) \\
=& \mathbb{E}\left[\|y^{l_k}\|_2^2\right] \prod_{l=l_k+1}^{L} (1+\alpha_l) \\
=& \mathbb{E}\left[\|y^{l_k-1}\|_2^2\right] \alpha_{l_k} \prod_{l=l_k+1}^{L} (1+\alpha_l) \\
=& \alpha_{l_k} \mathbb{E}[\|y^0\|_2^4] \prod_{\substack{l=1 \\ l\neq l_k}}^{L} (1+\alpha_l)
\end{aligned} \tag{61}$$

For the fourth moment, using the results of Props. 4 and 5 (taking into account that odd powers will vanish in expectation), it holds:

$$\begin{aligned}
\mathbb{E}\left[\|y^L\|_2^4\right] =& \mathbb{E}\left[\|y^{L-1}\|_2^4\right] + \alpha_L^2 \mathbb{E}\left[\|y^{L-1,m}\|_2^4\right] \\
&+ 4\alpha_L \mathbb{E}\left[\left({y^{L-1,m}}^\top y^{L-1}\right)^2\right] + 2\alpha_L \mathbb{E}\left[\|y^{L-1,m}\|_2^2 \cdot \|y^{L-1}\|_2^2\right]
\end{aligned} \tag{62}$$

Next, we analyze each term separately:

$$\mathbb{E}\left[\|y^{L-1,m}\|_2^4\right] = \mathbb{E}\left[\mathbb{E}[\|y^{L-1,m}\|_2^4 \mid R^{L-1}]\right] \tag{63}$$

Using the results of Props. 4 and 5:

$$\begin{aligned}
\mathbb{E}\left[\|y^{L-1,m}\|_2^4 \mid R^{L-1}\right] &= (1+2/n)\cdot(1+5/n)^{m-1} \cdot \|y^{L-1}\|_2^4 \\
&\approx (1+5/n)^m \cdot \|y^{L-1}\|_2^4
\end{aligned} \tag{64}$$

In addition,

$$\begin{aligned}
\mathbb{E}\left[\left({y^{L-1,m-1}}^\top y^{L-1}\right)^2\right] &= \frac{1}{n}\sum_{j_1,j_2,i_1,i_2} \mathbb{E}\left[y_{i_1}^{L-1,m-1} y_{i_2}^{L-1,m-1} y_{j_1}^{L-1} y_{j_2}^{L-1} w_{i_1,j_1}^{L,m} w_{i_2,j_2}^{L,m}\right] \\
&= \frac{1}{n}\mathbb{E}\left[\|y^{L-1,m-1}\|_2^2 \cdot \|y^{L-1}\|_2^2\right] \\
&= \frac{1}{n}\mathbb{E}\left[\|y^{L-1}\|_2^4\right]
\end{aligned} \tag{65}$$

and also,

$$\mathbb{E}\left[\|y^{L-1,m}\|_2^2 \cdot \|y^{L-1}\|_2^2\right] = \mathbb{E}\left[\|y^{L-1}\|_2^4\right] \tag{66}$$

Plugging it all into Eq. 62, by recursion, we have:

$$\mathbb{E}\left[\|y^L\|_2^4\right] \approx \mathbb{E}\left[\|y^{l_k}\|_2^4\right] \cdot \prod_{l=l_k+1}^{L} \beta_l \tag{67}$$

where,

$$\beta_l := 1 + 2\alpha_l\left(1+2/n\right) + \alpha_l^2\left(1+5/n\right)^m \tag{68}$$

In the reduced architecture, the transformation from layer $l_k - 1$ to layer l_k is given by an m layer fully connected network, with a linear layer on top, we can use the results from the vanilla case, and assigning $\|y^0\|_2^4 = 1$:

$$\begin{aligned}
\mathbb{E}\left[\|y^L\|_2^4\right] &= \alpha_{l_k}^2\left(1+2/n\right)\cdot(1+5/n)^{m-1} \prod_{l\neq l_k}^{L} \beta_l \\
&\approx \alpha_{l_k}^2\left(1+5/n\right)^m \prod_{l\neq l_k}^{L} \beta_l
\end{aligned} \tag{69}$$

Denoting $\rho = (1+5/n)^{\frac{m}{2}}$, and using the following:

$$\beta_l \approx (1+\alpha_l \rho)^2 \tag{70}$$

It follows that:

$$\begin{aligned}
\mathbb{E}[\mathcal{K}(x,x)^2] &\lessapprox \sum_{\mathbf{u},\mathbf{v}} \sqrt{\mathbb{E}[\|y^L_{(\mathbf{u})}\|_2^4]\mathbb{E}[\|y^L_{(\mathbf{v})}\|^2]} \\
&\approx (1+5/n)^m \sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v} \sqrt{\left(\prod_{l\neq l_u}^{L} \beta_l\right)\left(\prod_{l\neq l_v}^{L} \beta_l\right)} \\
&= (1+5/n)^m \sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v} \left[\prod_{l\neq l_u}(1+\rho\alpha_l)\right]\cdot\left[\prod_{l\neq l_v}(1+\rho\alpha_l)\right]
\end{aligned} \tag{71}$$

where $\mathbf{u} = (l_u, h_u)$ and $\mathbf{v} = (l_v, h_v)$.

Similarly, we have:

$$\mathbb{E}[\mathcal{K}(x,x)^2] \gtrapprox \sum_{\mathbf{k}} \mathbb{E}[\|J^{\mathbf{k}}\|_2^4] \approx (1+5/n)^m \cdot \sum_{\mathbf{u}} \alpha_{l_u}^2 \prod_{l\neq l_u}^{L} \beta_l = (1+5/n)^m \cdot \sum_{\mathbf{u}} \alpha_{l_u}^2 \prod_{l\neq l_u}^{L} (1+\rho\alpha_l)^2 \tag{72}$$

Using Eq. 61, we have that:

$$\mathbb{E}[\mathcal{K}(x,x)]^2 = \sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v} \left(\prod_{l\neq l_u}(1+\alpha_l)\right)\cdot\left(\prod_{l\neq l_v}(1+\alpha_l)\right) \tag{73}$$

This yields that:

$$\begin{aligned}
\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} &\lessapprox (1+5/n)^m \cdot \frac{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}\left(\prod_{l\neq l_u}(1+\rho\alpha_l)\right)\left(\prod_{l\neq l_v}(1+\rho\alpha_l)\right)}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}\left(\prod_{l\neq l_u}(1+\alpha_l)\right)\left(\prod_{l\neq l_v}(1+\alpha_l)\right)} \\
&\approx (1+5/n)^m \cdot \frac{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}\left(\prod_{l=1}^{L}(1+\rho\alpha_l)\right)\left(\prod_{l=1}^{L}(1+\rho\alpha_l)\right)}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}\left(\prod_{l=1}^{L}(1+\alpha_l)\right)\left(\prod_{l=1}^{L}(1+\alpha_l)\right)} \\
&= (1+5/n)^m \cdot \frac{\left(\prod_{l=1}^{L}(1+\rho\alpha_l)\right)^2}{\left(\prod_{l=1}^{L}(1+\alpha_l)\right)^2} \\
&= (1+5/n)^m \cdot \left(\prod_{l=1}^{L}\left(1+\frac{\alpha_l(\rho-1)}{1+\alpha_l}\right)\right)^2 \\
&\approx \exp\left[\frac{5m}{n} + \frac{C}{n}\sum_{l=1}^{L}\frac{\alpha_l}{1+\alpha_l}\right](1+\mathcal{O}(1/n))
\end{aligned} \tag{74}$$

For the lower bound, we have:

$$\begin{aligned}
\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} &\gtrapprox (1+5/n)^m \cdot \frac{\sum_{\mathbf{u}} \alpha_{l_u}^2 \left(\prod_{l\neq l_u}^{L}(1+\rho\alpha_l)\right)^2}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}\left(\prod_{l\neq l_u}(1+\alpha_l)\right)\left(\prod_{l\neq l_v}(1+\alpha_l)\right)} \\
&\approx \frac{\sum_{\mathbf{u}} \alpha_{l_u}^2}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}} \exp\left[\frac{5m}{n} + \frac{C}{n}\sum_{l=1}^{L}\frac{\alpha_l}{1+\alpha_l}\right](1+\mathcal{O}(1/n))
\end{aligned} \tag{75}$$

Since $\mathbb{E}[\mathcal{K}(x,x)^2] > \mathbb{E}[\mathcal{K}(x,x)]^2$, the lower bound is given by:

$$\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} \gtrapprox \max\left[1, \frac{\sum_{\mathbf{u}} \alpha_{l_u}^2}{\sum_{\mathbf{u},\mathbf{v}} \alpha_{l_u}\alpha_{l_v}} \exp\left[\frac{5m}{n} + \frac{C}{n}\sum_{l=1}^{L}\frac{\alpha_l}{1+\alpha_l}\right](1+\mathcal{O}(1/n))\right] \tag{76}$$

□

Theorem 4. *Let $f(x;w)$ be a constant width DenseNet (with $n'_0, n_l = n$ for all $l \in [L]$), with initialization constant $\alpha > 0$. Then, there exist constants $C_1, C_2 > 0$, such that:*

$$\max\left[1, \frac{C_1}{L\log(L)^2}\cdot\xi\right] \leq \eta(n,L) \leq \xi \tag{77}$$

where:

$$\xi = \exp\left[C_2/n\right]\cdot\left(1+\mathcal{O}(1/n)\right) \tag{78}$$

Proof. In the following proof, for the sake of notation simplicity, we omit the notation $\mathbf{k} = (l_k, h_k)$ in $y^l_{(\mathbf{k})}$, and assume that y^l stands for the reduced network $y^l_{(\mathbf{k})}$. The recursive formula for the intermediate outputs of the reduced network are given by:

$$y^l = \begin{cases} \sqrt{\frac{\alpha}{nl}}\sum_{h=k}^{l-1} W^{l,h}q^h & l_k < l \leq L \\ \sqrt{\frac{\alpha}{nl}}\sum_{h=0}^{l-1} W^{l,h}q^h & 1 \leq l < l_k \\ \sqrt{\frac{\alpha}{nl_k}} W^{l_k,h_k-1}q^{l_k-1} & l = l_k \end{cases} \tag{79}$$

with $q^h = \sqrt{2}\phi(y^h)$. We define, $\mu_l := \mathbb{E}\left[\|q^l\|_2^2\right]$. It follows that:

$$\mu_L = \mathbb{E}\left[\|q^L\|_2^2\right] = \frac{2\alpha}{Ln}\mathbb{E}\left[\left(\sum_{l=l_k}^{L-1} q^{l^\top}W^{L,l}\right)Z^L\left(\sum_{l=l_k}^{L-1} q^{l^\top}W^{L,l}\right)\right] = \frac{\alpha}{L}\sum_{l=l_k}^{L-1}\mu_l \tag{80}$$

where Z^l is a diagonal matrix holding in its diagonal the activation variables z^l_j for unit j in layer l.

Next, by telescoping the mean:

$$\begin{aligned} \mu_L &= \frac{\alpha}{L}\sum_{l=l_k}^{L-1}\mu_l = \frac{\alpha\mu_{L-1}}{L} + \frac{L-1}{L}\mu_{L-1} = \mu_{L-1}\left(1+\frac{\alpha-1}{L}\right) \\ &= \mu_{l_k+1}\prod_{l=l_k+2}^{L}\left(1+\frac{\alpha-1}{l}\right) = \frac{\alpha}{l_k+1}\mu_{l_k}\prod_{l=l_k+2}^{L}\left(1+\frac{\alpha-1}{l}\right) \\ &= \frac{\alpha}{l_k+1}\mu_0\prod_{\substack{l=1 \\ l\neq l_k+1}}^{L}\left(1+\frac{\alpha-1}{l}\right) \approx \frac{\alpha}{l_k+1}\prod_{l=1}^{L}\left(1+\frac{\alpha-1}{l}\right) \end{aligned} \tag{81}$$

and so:

$$\mathbb{E}[\mathcal{K}(x,x)]^2 = \left(\sum_{l_k=1}^{L}\mu_L\right)^2 \approx \left(\sum_{l_k=1}^{L}\frac{\alpha}{l_k+1}\right)^2\prod_{l=1}^{L}\left(1+\frac{\alpha-1}{l}\right)^2 \approx \alpha^2\log(L)^2\prod_{l=1}^{L}\left(1+\frac{\alpha-1}{l}\right)^2 \tag{82}$$

For the fourth moment:

$$\begin{aligned} &\mathbb{E}\left[\|q^L\|_2^4\right] \\ =&\frac{4\alpha^2}{n^2L^2}\mathbb{E}\left[\left(\sum_{l=l_k}^{L-1}\left(q^{l^\top}W^{L,l}\right)Z^L\sum_{l=l_k}^{L-1}\left(q^{l^\top}W^{L,l}\right)\right)^2\right] \\ =&\frac{4\alpha^2}{n^2L^2}\mathbb{E}\left[\left(\sum_{l_1=l_k}^{L-1}\left(q^{l_1\top}W^{L,l_1}\right)Z^L\sum_{l_2=l_k}^{L-1}\left(q^{l_2\top}W^{L,l_2}\right)\sum_{l_3=l_k}^{L-1}\left(q^{l_3\top}W^{L,l_3}\right)Z^L\sum_{l_4=l_k}^{L-1}\left(q^{l_4\top}W^{L,l_4}\right)\right)\right] \end{aligned} \tag{83}$$

We denote:

$$C_{l,l'} = \mathbb{E}\left[\|q^l\|_2^2\cdot\|q^{l'}\|_2^2\right] \tag{84}$$

Using the results from the vanilla architecture, we have:

$$C_{L,L} = \frac{\alpha^2(n+5)}{nL^2} \sum_{l_1,l_2=l_k}^{L-1} C_{l_1,l_2} \tag{85}$$

From Eq. 85, it also holds that:

$$\sum_{l_1,l_2=l_k}^{L-2} C_{l_1,l_2} = \frac{n(L-1)^2}{\alpha^2(n+5)} \cdot C_{L-1,L-1} \tag{86}$$

It then follows:

$$\begin{aligned}
\mathbb{E}\left[\|q^L\|_2^4\right] =& C_{L,L} \\
=& \frac{\alpha^2(1+5/n)}{L^2} \sum_{l_1,l_2=l_k}^{L-1} C_{l_1 l_2} \\
=& \frac{\alpha^2(1+5/n)}{L^2} \left(C_{L-1,L-1} + \sum_{l_1,l_2=l_k}^{L-2} C_{l_1 l_2} + 2\sum_{l=l_k}^{L-2} C_{L-1,l} \right) \\
=& \frac{\alpha^2(1+5/n)}{L^2} \left(C_{L-1,L-1} + \frac{(L-1)^2 n}{\alpha^2(n+5)} C_{L-1,L-1} + 2\sum_{l=l_k}^{L-2} C_{L-1,l} \right)
\end{aligned} \tag{87}$$

The following also holds for all $l_1 > l_2 \geq l_k$:

$$C_{l_1,l_2} = \frac{\alpha}{nl_1} \mathbb{E}\left[(\sum_{l=l_k}^{l_1-1} q^{l\top} W^{l_1,l} Z^{l_1})^2 \|q^{l_2}\|_2^2 \right] = \frac{\alpha}{l_1} \sum_{l=l_k}^{l_1-1} C_{l,l_2} \tag{88}$$

and so:

$$\begin{aligned}
C_{L,L} &= \frac{\alpha^2(n+5)}{nL^2} \left(C_{L-1,L-1} + \frac{(L-1)^2 n}{\alpha^2(n+5)} C_{L-1,L-1} + \frac{2\alpha}{L-1} \sum_{l_1=l_k}^{L-2} \sum_{l_2=l_k}^{L-2} C_{l_1,l_2} \right) \\
&= \frac{\alpha^2(n+5)}{nL^2} \left(C_{L-1,L-1} + \frac{(L-1)^2 n}{\alpha^2(n+5)} C_{L-1,L-1} + \frac{2n(L-1)}{\alpha(n+5)} C_{L-1,L-1} \right) \\
&= \frac{\alpha^2(n+5)}{nL^2} C_{L-1,L-1} \left(1 + \frac{(L-1)^2 n}{\alpha^2(n+5)} + \frac{2n(L-1)}{\alpha(n+5)} \right) \\
&= C_{L-1,L-1} \left(\left(1 + \frac{\alpha-1}{L}\right)^2 + \frac{5\alpha^2}{nL^2} \right)
\end{aligned} \tag{89}$$

Recursively, we have:

$$C_{L,L} = C_{l_k+1,l_k+1} \prod_{l=l_k+2}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2} \right) \tag{90}$$

For the reduced architecture, the transition from q^{l_k} to q^{l_k+1} is a vanilla ReLU block, and so using the result from the vanilla architecture:

$$\begin{aligned}
C_{L,L} &= C_{l_k,l_k} \frac{\alpha^2(n+5)}{n(l_k+1)^2} \prod_{l=l_k+2}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2} \right) \\
&= \frac{\alpha^2(n+5)}{n(l_k+1)^2} \prod_{l \neq l_k+1} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2} \right) \\
&\approx \frac{\alpha^2(n+5)}{n(l_k+1)^2} \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2} \right)
\end{aligned} \tag{91}$$

where we assigned $C_{0,0} = 1$. It follows:

$$\begin{aligned}
\mathbb{E}[\mathcal{K}(x,x)^2] &\lessapprox \sum_{\mathbf{u},\mathbf{v}} \sqrt{\mathbb{E}\left[\|y^L_{(\mathbf{u})}\|_2^4\right] \cdot \mathbb{E}\left[\|y^L_{(\mathbf{v})}\|_2^2\right]} \\
&\approx \left(\sum_{l_k=1}^{L} \frac{1}{l_k+1}\right)^2 \cdot \frac{\alpha^2(n+5)}{n} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right) \\
&\approx \log(L)^2 \cdot \frac{\alpha^2(n+5)}{n} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right)
\end{aligned} \tag{92}$$

Similarly, we have:

$$\begin{aligned}
\mathbb{E}[\mathcal{K}(x,x)^2] &\gtrapprox \sum_{l_k} \mathbb{E}[\|J^{\mathbf{k}}\|_2^4] \\
&= \sum_{l_k=1}^{L} \frac{\alpha^2(n+5)}{n(l_k+1)^2} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right) \\
&\approx \frac{\alpha^2(n+5)}{nL} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right)
\end{aligned} \tag{93}$$

This yields that:

$$\begin{aligned}
\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} &\lessapprox \frac{\frac{n+5}{n} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right)}{\prod_{l=1}^{L} \left(1 + \frac{\alpha-1}{l}\right)^2} \\
&= \frac{n+5}{n} \cdot \prod_{l=1}^{L} \left(1 + \frac{5\alpha^2}{n(l+\alpha-1)^2}\right) \\
&\approx \exp\left[\sum_{l=1}^{L} \frac{5\alpha^2}{n(l+\alpha-1)^2}\right] \cdot (1 + \mathcal{O}(1/n)) \\
&\approx \exp\left[C/n\right] \cdot (1 + \mathcal{O}(1/n))
\end{aligned} \tag{94}$$

For the lower bound, we have:

$$\begin{aligned}
\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} &\gtrapprox \frac{\frac{n+5}{n} \cdot \prod_{l=1}^{L} \left(\left(1 + \frac{\alpha-1}{l}\right)^2 + \frac{5\alpha^2}{nl^2}\right)}{L\log(L)^2 \prod_{l=1}^{L} \left(1 + \frac{\alpha-1}{l}\right)^2} \\
&\approx \frac{1}{L\log(L)^2} \cdot \exp\left[C/n\right] \cdot (1 + \mathcal{O}(1/n))
\end{aligned} \tag{95}$$

Since $\mathbb{E}[\mathcal{K}(x,x)^2] > \mathbb{E}[\mathcal{K}(x,x)]^2$, the lower bound is given by:

$$\frac{\mathbb{E}[\mathcal{K}(x,x)^2]}{\mathbb{E}[\mathcal{K}(x,x)]^2} \gtrapprox \max\left[1, \frac{1}{L\log(L)^2} \cdot \exp\left[C/n\right] \cdot (1 + \mathcal{O}(1/n))\right] \tag{96}$$

□