STATISTICAL CHARACTERIZATION OF DEEP NEURAL NETWORKS AND THEIR SENSITIVITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their ubiquity, it remains an active area of research to fully understand deep neural networks (DNNs) and the reasons of their empirical success. We contribute to this effort by introducing a principled approach to statistically characterize DNNs and their sensitivity. By distinguishing between randomness from input data and from model parameters, we study how central and non-central moments of network activation and sensitivity evolve during propagation. Thereby, we provide novel statistical insights on the hypothesis space of input-output mappings encoded by different architectures. Our approach applies both to fully-connected and convolutional networks and incorporates most ingredients of modern DNNs: rectified linear unit (ReLU) activation, batch normalization, skip connections.[1]

1 INTRODUCTION

While the empirical success of deep neural networks is not disputed anymore, a full understanding of these models is not yet achieved (Zhang et al., 2016; Dinh et al., 2017; Neyshabur et al., 2017). As no exception to this rule, advances in the design of neural network architectures have more often come from the relentless race of practical applications rather than by principled approaches. Consequently many common practices and rules of thumb are still awaiting for theoretical validation.

An important obstacle in the characterization of neural networks is the complex interplay of different sources of randomness. In that respect, despite winning successes both theoretically (Poole et al., 2016; Schoenholz et al., 2016; Yang & Schoenholz, 2017; Pennington et al., 2018) and empirically (Pennington et al., 2017; Xiao et al., 2018), the mean-field theory of neural networks fails to distinguish between the randomness from input data and model parameters. As a result, input data is only modeled in the rudimentary case of two correlated signals with Gaussian pre-activation distribution. In Balduzzi et al. (2017) input data is similarly modeled using two correlated signals with typical activation patterns. Another path of research considers input data as a 1-dimensional manifold of evolving length and curvature (Poole et al., 2016; Raghu et al., 2017). All cases are limited in their scope or simplifying assumptions.

In this paper, we introduce a novel approach to statistically characterize deep neural networks and their sensitivity. Only mild assumptions are required and the usual simplifications of infinite width, gaussianity or typical activation patterns are not made. Both fully-connected and convolutional networks are encompassed and the commonly used techniques of batch normalization and skip connections are incorporated. The key of our methodology is to consider statistical moments with respect to input data as random variables which depend on model parameters. By studying how different architecture choices influence the evolution with depth of these moments, we provide statistical insights on the corresponding hypothesis spaces of input-output mappings. Our findings span the topics of pseudo-linearity, exploding sensitivity, exponential and power-law evolution with depth.

2 PROPAGATION

We start by formulating the propagation for neural networks with neither batch normalization nor skip connections, that we refer as *vanilla networks*. The formulation will be slightly adapted in section 6 with *batch-normalized feedforward nets*, and in section 7 with *batch-normalized resnets*.

[1] Code to reproduce all results will be made available upon publication.

Clean propagation. Suppose that we are given a random tensorial input $\mathbf{x} \in \mathbb{R}^{n\times\cdots\times n\times N_0}$ which is spatially d-dimensional with spatial extent of n in each direction and N_0 channels. We further suppose that this input is not trivially zero such that $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^2_{\boldsymbol{\alpha},\mathrm{c}}] > 0$,[2] where $\boldsymbol{\alpha}$ denotes the spatial position and c the channel. This input is fed into a d-dimensional convolutional neural network with periodic boundary conditions and constant spatial extent n.[3] For each layer, we denote N_l the number of channels or *width*, K_l the convolutional spatial extent, $\boldsymbol{\omega}^l \in \mathbb{R}^{K_l\times\cdots\times K_l\times N_{l-1}\times N_l}$ the weight tensors, $\mathbf{b}^l \in \mathbb{R}^{N_l}$ the biases, and $\mathbf{x}^l, \mathbf{y}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ the tensors of post-activations and pre-activations. We further denote ϕ the activation function and we adopt the convention $\mathbf{x}^0 \equiv \mathbf{x}$. The propagation at each layer l is given by $\mathbf{x}^l = \phi(\boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l)$, where $\boldsymbol{\beta}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ is the tensor with repeated version of $\mathbf{b}^l$ at each spatial position. From now on, we refer to the propagated tensor $\mathbf{x}^l$ as the *signal*.

Noisy propagation. Next we suppose that the input signal $\mathbf{x}$ is corrupted by a small *white noise* tensor $\boldsymbol{\epsilon} \in \mathbb{R}^{n\times\cdots\times n\times N_0}$ with independent and identically distributed components such that $\mathbb{E}_{\boldsymbol{\epsilon}}[\boldsymbol{\epsilon}_i\boldsymbol{\epsilon}_j] = \sigma^2_{\boldsymbol{\epsilon}}\delta_{ij}$ ($\sigma_{\boldsymbol{\epsilon}} \ll 1$), with δ_{ij} the Kronecker delta. The noisy signal is propagated in the same neural network and we keep track of the noise corruption with the tensor $\boldsymbol{\epsilon}^l \equiv \Phi^l(\mathbf{x}+\boldsymbol{\epsilon}) - \Phi^l(\mathbf{x})$, where Φ^l is the input-output mapping $\mathbf{x}^l = \Phi^l(\mathbf{x})$. Again with the convention $\boldsymbol{\epsilon}^0 \equiv \boldsymbol{\epsilon}$, the simultaneous propagation of the *clean signal* $\mathbf{x}^l$ and the *noise* $\boldsymbol{\epsilon}^l$ is given by

$$\mathbf{y}^l = \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l, \quad \mathbf{x}^l = \phi(\mathbf{y}^l), \tag{1}$$

$$\boldsymbol{\epsilon}^l = \boldsymbol{\omega}^l * \boldsymbol{\epsilon}^{l-1} \odot \phi'(\mathbf{y}^l), \tag{2}$$

where $\odot$ denotes the element-wise tensor multiplication and Eq. (2) is obtained by taking the derivative in Eq. (1). As shown by Eq. (2), for given $\mathbf{x}$ the mapping from $\boldsymbol{\epsilon}$ to $\boldsymbol{\epsilon}^l$ is linear. The noise $\boldsymbol{\epsilon}^l$ thus stays centered with respect to $\boldsymbol{\epsilon}$ during propagation with $\forall\boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\boldsymbol{\epsilon}^l_{\boldsymbol{\alpha},\mathrm{c}}\right] = 0$.

To get rid of the dependence on $\sigma_{\boldsymbol{\epsilon}}$, we introduce the random *sensitivity tensor* as the rescaling of the noise with unit initial variance: $\mathbf{s}^0 \equiv \mathbf{s} \equiv \boldsymbol{\epsilon} \,/\, \sigma_{\boldsymbol{\epsilon}}$ and $\mathbf{s}^l \equiv \boldsymbol{\epsilon}^l \,/\, \sigma_{\boldsymbol{\epsilon}}$. Due to the linearity of Eq. (2), the sensitivity tensor $\mathbf{s}^l$ is the result of the simultaneous propagation of $\mathbf{x}^l$ and $\mathbf{s}^l$ in Eq. (1) and (2). We also have $\mathbb{E}_{\mathbf{s}}[\mathbf{s}_i\mathbf{s}_j] = \delta_{ij}$ and $\forall\boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\mathbf{s}}\left[\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}}\right] = 0$. The sensitivity tensor encodes *derivative information* while avoiding the burden of increased dimensionality (see computation in Appendix D.1 for an illustration). This will prove very useful.

Further scope. We restrict our analysis to the ReLU *activation function*: $\phi(\mathbf{y}^l) = \max(\mathbf{y}^l, 0)$. This is partly due to lack of space and partly because ReLU networks are the most widely used in practice. Note however that the formulation with convolutional neural networks does not exclude fully-connected networks, obtained simply as a subcase with $n = 1$.

3 Input data randomness – moments, normalized sensitivity

3.1 Input data randomness

To understand the importance of the input data distributions $P_{\mathbf{x}}(\mathbf{x}^l)$ and $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$, let us adopt a geometrical perspective on Eq. (1) and Eq. (2) in the context of *classification*. The neural network is fed a noisy point cloud with different classes spread throughout space. Its goal is layer after layer to modify this point cloud in $\mathbb{R}^{n\times\cdots\times n\times N_l}$ in order to better group points from the same class and separate points from different classes. This continues until the point cloud reaches the final linear separation. The evolution of the point cloud and its derivative across layers, $P_{\mathbf{x}}(\mathbf{x}^l)$ and $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$, is of crucial importance since it characterizes the internal neural network machinery.

Note the distinction between $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$ on one side, and $P_{\mathbf{x},\boldsymbol{\omega},\boldsymbol{\beta}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s},\boldsymbol{\omega},\boldsymbol{\beta}}(\mathbf{s}^l)$ on the other side. As an illustration, consider a neural network which has shrunk its input $\mathbf{x}$ to a point mass distribution $P_{\mathbf{x}}(\mathbf{x}^{l-1}) = \delta_{p_{l-1}}$. For given $\boldsymbol{\omega}^l$ and $\boldsymbol{\beta}^l$, the propagation of Eq. (1) maps this distribution to another point mass $P_{\mathbf{x}}(\mathbf{x}^l) = \delta_{p_l}$. On the other side, $P_{\mathbf{x},\boldsymbol{\omega},\boldsymbol{\beta}}(\mathbf{x}^l)$ has density spreading in the whole ambient space $\mathbb{R}^{n\times\cdots\times n\times N_l}$ and misses the *distributional pathology*.

[2] Whenever $\boldsymbol{\alpha}$ and c are considered as random variables they are supposed uniformly sampled among all spatial positions $\{1,\ldots,n\}^d$ and all channels $\{1,\ldots,N_l\}$.

[3] The assumptions of periodic boundary conditions and constant spatial extent n are made for simplicity of the analysis. Possible relaxations are discussed in section C.2.

3.2 MOMENTS

In order to keep track of $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$, our next challenge is the tensorial structure of $\mathbf{x}^l$ and $\mathbf{s}^l$. In a similar way as batch normalization, we consider feature maps at different spatial positions as interwoven *sub-signals* of the tensorial *meta-signal*. Statistically, we work at the granularity of sub-signals and we treat equally the randomness of the input tensors $\mathbf{x}$, $\mathbf{s}$ and the spatial position $\boldsymbol{\alpha}$. This brings the definition of the *feature map vector* and *centered feature map vector* of $\mathbf{x}^l$ as

$$\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha}) \equiv \mathbf{x}^l_{\boldsymbol{\alpha},:}, \qquad \hat{\mathbf{f}}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha}) \equiv \mathbf{x}^l_{\boldsymbol{\alpha},:} - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{x}^l_{\boldsymbol{\alpha},:}\right],$$

where $\boldsymbol{\alpha}$ is uniformly sampled in $\{1, \dots, n\}^d$ and the denotation $\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})$ reminds the randomness of both $\mathbf{x}^l$ and $\boldsymbol{\alpha}$. For any order p, the *non-central moment* and *central moment* of $\mathbf{x}^l$ per-channel and averaged over channels are defined as

$$\nu_{p,\text{c}}(\mathbf{x}^l) \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})^p_{\text{c}}\right], \qquad \mu_{p,\text{c}}(\mathbf{x}^l) \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\hat{\mathbf{f}}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})^p_{\text{c}}\right],$$
$$\nu_{p}(\mathbf{x}^l) \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\text{c}}\left[\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})^p_{\text{c}}\right], \qquad \mu_{p}(\mathbf{x}^l) \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\text{c}}\left[\hat{\mathbf{f}}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})^p_{\text{c}}\right].$$

The previous definitions are further extended to any random tensor.

3.3 NORMALIZED SENSITIVITY

Normalized sensitivity. Using the moments of $\mathbf{x}^l$ and $\mathbf{s}^l$ we finally define our key metric for the statistical characterization of neural networks that we refer as the *normalized sensitivity* ζ^l:

$$\zeta^l \equiv \left(\frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)}\right)^{1/2}, \tag{3}$$

where $\mu_2(\mathbf{s}^l)$ measures *sensitivity* and $\mu_2(\mathbf{x}^l)$, $\mu_2(\mathbf{x}^0)$ measure *signal informativeness*. Again in the classification task where the goal is to set apart different signals, informativeness is measured by $\mu_2(\mathbf{x}^l)$ since a constant shift applied to all signals is uninformative. This is summarized by the property of ζ^l to measure an expected sensitivity when neural network input and output signals are rescaled to have unit variances (proof and visual illustration in Appendix D.2).

Normalized Sensitivity and Signal-to-Noise. Now let us push further the view of noisy propagation with the terminology of *signal-to-noise ratio* SNR^l and *noise factor* F^l:

$$SNR^l \equiv \left(\frac{\sigma^l_{\text{signal}}}{\sigma^l_{\text{noise}}}\right)^2 = \frac{\mu_2(\mathbf{x}^l)}{\mu_2(\boldsymbol{\epsilon}^l)}, \qquad F^l \equiv \frac{SNR^0}{SNR^l} = \frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)} = (\zeta^l)^2,$$

where we used the definitions of $\mu_2(\mathbf{x}^l)$, $\mu_2(\boldsymbol{\epsilon}^l)$ and $\mu_2(\boldsymbol{\epsilon}^l) = \sigma^2_{\boldsymbol{\epsilon}}\mu_2(\mathbf{s}^l)$ and $\mu_2(\boldsymbol{\epsilon}^0) = \sigma^2_{\boldsymbol{\epsilon}}$. In logarithmic decibel scale, $SNR^l_{\text{dB}} = SNR^0_{\text{dB}} - 20\log_{10}\zeta^l$. The normalized sensitivity ζ^l then directly measures how the neural network degrades ($\zeta^l > 1$) or enhances ($\zeta^l < 1$) the input signal-to-noise ratio. The *factor of noise equivalence* means that neural networks with high ζ^l are essentially *noise amplifiers*.

Normalized Sensitivity and Generalization. The relevance of ζ^l is further supported by several studies relating sensitivity and generalization for fully connected networks (Sokolic et al., 2017; Novak et al., 2018; Philipp & Carbonell, 2018). Notably the coefficient defined in Philipp & Carbonell (2018) is equivalent to the normalized sensitivity ζ^l in the fully-connected case (details on equivalence and reasons for our change of terminology in Appendix D.1).

4 MODEL PARAMETERS RANDOMNESS

We now introduce model parameters as the second source of randomness. We only consider *untrained networks* at initialization. Due to the randomness of the initialization point and its independence from input data distribution, this can be seen as characterizing the hypothesis space of input-output mappings encoded by the architecture. We assume that *initialization is standard*: (i) weights and biases are initialized following He et al. (2015), (ii) when pre-activations are

batch-normalized, scale and shift batch normalization parameters are initialized with ones and zeros respectively.

Our methodology from now on is to consider all moment-related quantities such as $\mu_p(\mathbf{x}^l)$, $\mu_p(\mathbf{s}^l)$, $\nu_p(\mathbf{x}^l)$, $\nu_p(\mathbf{s}^l)$, ζ^l as random variables depending on $(\boldsymbol{W}^1, \mathbf{b}^1, \dots, \boldsymbol{W}^l, \mathbf{b}^l)$. We introduce the notation $\Theta^l = (\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \dots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ for the full set of parameters and the notation $\theta^l = \Theta^l | \Theta^{l-1}$ for the conditional set of parameters, when $(\boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ are considered as random and $(\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \boldsymbol{\omega}^{l-1}, \boldsymbol{\beta}^{l-1})$ as given. Furthermore we denote $\delta\mu_2^l(\mathbf{x})$ the geometric increments such that $\log \delta\mu_2(\mathbf{x}^l) = \log \mu_2(\mathbf{x}^l) - \log \mu_2(\mathbf{x}^{l-1})$.

Evolution with Depth. We are now able to write the evolution with depth of $\mu_2(\mathbf{x}^l)$ as

$$\log \mu_2(\mathbf{x}^l) - \log \mu_2(\mathbf{x}^0) = \sum\nolimits_k \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)] + \sum\nolimits_k \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)] + \sum\nolimits_k \log \delta\mu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)], \tag{4}$$

where Eq. (4) is obtained using $\log \mu_2(\mathbf{x}^l) - \log \mu_2(\mathbf{x}^0) = \sum_k \log \delta\mu_2(\mathbf{x}^k)$ and by decomposing each term in the sum with telescoping terms. Let us define $\delta_r\mu_2(\mathbf{x}^k)$ such that $\delta\mu_2(\mathbf{x}^k) = \delta_r\mu_2(\mathbf{x}^k)\, \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]$, and denote $\overline{m}$, $\underline{m}$, $\underline{s}$ the three different terms in Eq. (4):

$$\overline{m}[\mu_2(\mathbf{x}^k)] \equiv \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)], \tag{5}$$

$$\underline{m}[\mu_2(\mathbf{x}^k)] \equiv \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)] = \mathbb{E}_{\theta^k}[\log \delta_r\mu_2(\mathbf{x}^k)], \tag{6}$$

$$\underline{s}[\mu_2(\mathbf{x}^k)] \equiv \log \delta\mu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)] = \log \delta_r\mu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \delta_r\mu_2(\mathbf{x}^k)], \tag{7}$$

where Eq. (6) is obtained by entering $\log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]$ into the conditional expectation and the logarithm, and Eq. (7) is obtained using $\mathbb{E}_{\theta^k}[\,\mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)]] = \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)]$.

Discussion. First we note that $\overline{m}[\mu_2(\mathbf{x}^k)]$ and $\underline{m}[\mu_2(\mathbf{x}^k)]$ are random variables which depend on Θ^{k-1} while $\underline{s}[\mu_2(\mathbf{x}^k)]$ is a random variable which depends on Θ^k. We also note that $\underline{m}[\mu_2(\mathbf{x}^k)] < 0$ by log-concavity and that $\underline{s}[\mu_2(\mathbf{x}^k)]$ is centered: $\mathbb{E}_{\Theta^k}[\underline{s}[\mu_2(\mathbf{x}^k)]] = 0$.

Under standard initialization, each channel provides an independent contribution to $\mu_2(\mathbf{x}^k) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^k, \boldsymbol{\alpha})_{\mathrm{c}}^2]$. As a consequence, for large N_k the relative increment $\delta_r\mu_2(\mathbf{x}^k)$ has low expected deviation to 1, meaning with high probability that $|\log \delta_r\mu_2(\mathbf{x}^k)| \ll 1$, $|\underline{m}[\mu_2(\mathbf{x}^k)]| \ll 1$, $|\underline{s}[\mu_2(\mathbf{x}^k)]| \ll 1$. In addition, $\underline{s}[\mu_2(\mathbf{x}^k)]$ is centered and *non-correlated* at different k so its sum scales as $\sqrt{l}$, whereas the sums of $\overline{m}[\mu_2(\mathbf{x}^k)]$ and $\underline{m}[\mu_2(\mathbf{x}^k)]$ scale as l (see Lemma 10 in Appendix E.1). The term $\underline{s}[\mu_2(\mathbf{x}^k)]$ is thus doubly negligible. In summary, the evolution with depth is dominated by $\overline{m}[\mu_2(\mathbf{x}^k)]$ when this term is non-vanishing and by $\underline{m}[\mu_2(\mathbf{x}^k)]$ otherwise. The exact same analysis can be applied to $\nu_2(\mathbf{x}^l)$ and to sensitivity moments $\nu_2(\mathbf{s}^l)$, $\mu_2(\mathbf{s}^l)$. It can also be applied to the ratio $\mu_2(\mathbf{s}^l) \,/\, \mu_2(\mathbf{x}^l)$ and thus to ζ^l.

Further notation. From now on, the geometric increment of any quantity is denoted with δ. The definitions of $\overline{m}$, $\overline{m}$ and $\underline{s}$ in Eq. (5), (6) and (7) are extended to other central and non-central moments of signal and sensitivity as well as ζ^l with $\overline{m}[\zeta^l] = \frac{1}{2}\big(\overline{m}[\mu_2(\mathbf{s}^l)] - \overline{m}[\mu_2(\mathbf{x}^l)]\big)$, $\underline{m}[\zeta^l] = \frac{1}{2}\big(\underline{m}[\mu_2(\mathbf{s}^l)] - \underline{m}[\mu_2(\mathbf{x}^l)]\big)$, $\underline{s}[\zeta^l] = \frac{1}{2}\big(\underline{s}[\mu_2(\mathbf{s}^l)] - \underline{s}[\mu_2(\mathbf{x}^l)]\big)$.

The approximation of dominating terms such as $\overline{m}[\mu_2(\mathbf{x}^k)]$ in Eq. (4) is denoted with $\simeq$. To make it precise, we write $a \simeq b$ when $a(1+\delta_a) = b(1+\delta_b)$ with $|\delta_a| \ll 1$, $|\delta_b| \ll 1$ with high probability. We write $a \lesssim b$ when $a(1+\delta_a) \leq b(1+\delta_b)$ with $|\delta_a| \ll 1$, $|\delta_b| \ll 1$ with high probability. From now on we further assume that the *width is large*. We stress that this assumption is milder than mean-field since dominating terms such as $\overline{m}[\mu_2(\mathbf{x}^k)]$ in Eq. (4) remain random variables and since we do not require the signal $\mathbf{x}^l$ to be Gaussian.

5 Vanilla Networks

We are now fully equipped to statistically characterize neural network architectures. We start by analyzing vanilla networks corresponding to the equations of propagation introduced in section 2.

Theorem 1. Moments of vanilla networks. (proof in Appendix E.2) *Denote A_l the event $\{\nu_2(\mathbf{x}^l) > 0\}$ and A_l' the complementary event $\{\nu_2(\mathbf{x}^l) = 0\} = \{\mathbb{P}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0] = 1\}$. Then:*

(i) $\prod_{k=1}^{l}\left(1-2^{-N_k}\right) \leq \mathbb{P}[A_l] \leq \prod_{k=1}^{l}\left(1-2^{-K_k^d N_{k-1} N_k}\right)$

(ii) There exist positive constants m_{min}, m_{max}, v_{min}, $v_{\text{max}} > 0$ *and sequences of random variables* $(m_l), (m'_l), (s_l), (s'_l)$ *such that under* A_l, s_l, s'_l *are centered and*

$$\log \nu_2(\mathbf{x}^l) = -l m_l + \sqrt{l} s_l + \log \nu_2(\mathbf{x}^0), \quad m_{\text{min}} \leq m_l \leq m_{\text{max}}, \quad v_{\text{min}} \leq \text{Var}_{\Theta^l | A_l}[s_l] \leq v_{\text{max}},$$
$$\log \mu_2(\mathbf{s}^l) = -l m'_l + \sqrt{l} s'_l, \quad m_{\text{min}} \leq m'_l \leq m_{\text{max}}, \quad v_{\text{min}} \leq \text{Var}_{\Theta^l | A_l}[s'_l] \leq v_{\text{max}}.$$

Discussion. (i) is related to the collapse of ReLU networks, which is studied in Lu et al. (2018). Any vanilla ReLU network almost surely collapses to 0 but the number of required layer is exponential in the width N_l. In practice, it is not a real problem.

(ii) implies that moments of $\mathbf{x}^l$ and $\mathbf{s}^l$ can be written $\nu_2(\mathbf{x}^l) = \nu_2(\mathbf{x}^0)\exp(-l m_l + \sqrt{l} s_l)$ and $\mu_2(\mathbf{s}^l) = \exp(-l m'_l + \sqrt{l} s'_l)$. This behaviour comes from the particularity of He et al. (2015) to keep stable $\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\text{c}}[(\mathbf{x}^l_{\boldsymbol{\alpha},\text{c}})^2] = \mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$ and $\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\text{c}}[(\mathbf{s}^l_{\boldsymbol{\alpha},\text{c}})^2] = \mathbb{E}_{\Theta^l}[\mu_2(\mathbf{s}^l)]$ during propagation, which results in vanishing $\overline{m}[\nu_2(\mathbf{x}^l)]$ and $\overline{m}[\mu_2(\mathbf{s}^l)]$. The dominating terms in Eq. (4) are then $\underline{m}[\nu_2(\mathbf{x}^k)] < 0$ and $\underline{m}[\mu_2(\mathbf{s}^k)] < 0$ (see Appendix E.3 for details). The small negative drift and the increasing variance of $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$ is clear in Fig. 1d and Fig. 1e. It is also clear that the distribution of $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$ is nearly Gaussian and therefore the distribution of $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ is nearly lognormal. Since a lognormal distribution $\exp(\mu + \sigma X)$ with $X \sim \mathcal{N}(0,1)$ has expectation equal to $\exp(\mu + \sigma^2/2)$, the increasing variance of $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$ must be compensated by small negative drift in order for their exponential $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ to have stable expectations $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$ and $\mathbb{E}_{\Theta^l}[\mu_2(\mathbf{s}^l)]$. Note that the diffusion happens in log-space since layer composition amounts to multiplicative random effect in real space.

As a consequence of (ii) and Chebyshev's inequality, conditionally on A_l the variables $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ still converge in probability to 0 (proof in Appendix E.4). He et al. (2015) thus manages to stabilize expectations with respect to all realizations Θ^l. However in practice we only see a single realization Θ^l and for large l this leads with high probability to vanishing network signal and sensitivity (i.e. activations and gradients). Note that this is a finite-width effect and the terms $\underline{m}[\nu_2(\mathbf{x}^l)], \underline{m}[\mu_2(\mathbf{s}^l)], \underline{s}[\nu_2(\mathbf{x}^l)], \underline{s}[\mu_2(\mathbf{s}^l)]$ also vanish in the limit of infinite width.

Theorem 2. Normalized Sensitivity increments of vanilla networks. (proof in Appendix F.1) *Under* A_{l-1}, *the dominating term in the evolution of the normalized sensitivity is:*

$$\delta\zeta^l \simeq \exp\left(\overline{m}_{\text{vanilla}}\left[\zeta^l\right]\right) = \left(1 - \mathbb{E}_{\text{c},\theta^l | A_{l-1}}\left[\frac{\nu_{1,\text{c}}(\mathbf{y}^{l,+})\nu_{1,\text{c}}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-1/2}, \tag{8}$$

where $\mathbf{y}^{l,+} = \max(\mathbf{y}^l, 0)$ *and* $\mathbf{y}^{l,-} = \max(-\mathbf{y}^l, 0)$.

Discussion. An immediate consequence of Theorem 2 is that $\delta\zeta^l \gtrsim 1$, meaning that normalized sensitivity always increases with depth for ReLU vanilla networks. To further understand the behaviour of ζ^l we proceed by contradiction. Suppose that there is an event D with probability $\mathbb{P}[D] > 0$ under which $\log \zeta^l$ has a drift larger than the diffusion. Under D the ratio $\mu_2(\mathbf{x}^l) \,/\, \nu_2(\mathbf{x}^l)$ then converges in probability to 0 and the variance $\mu_2(\mathbf{x}^l)$ becomes arbitrary smaller than the average magnitude $\nu_2(\mathbf{x}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[||\mathbf{x}^l||_2^2] \,/\, N_l$ (proof in Appendix F.2). All inputs $\mathbf{x}$ are then mapped to a very localized region and the distribution of $\mathbf{x}^l$ resembles that of a single point. In turn, this implies that with high probability a given pre-activation channel will have all its values concentrated around either a single positive or a single negative value. This means that with high probability either $\mathbf{y}^{l,+}_{:,\text{c}} = 0$ for nearly all $\mathbf{x}, \boldsymbol{\alpha}$, or $\mathbf{y}^{l,-}_{:,\text{c}} = 0$ for nearly all $\mathbf{x}, \boldsymbol{\alpha}$, and thus $\nu_{1,\text{c}}(\mathbf{y}^{l,+})\nu_{1,\text{c}}(\mathbf{y}^{l,-}) \,/\, \mu_2(\mathbf{x}^{l-1}) \ll 1$ and $\delta\zeta^l \simeq 1$. This contradicts the presence of the drift larger than the diffusion in $\log \zeta^l$.

The previous argument shows that $\delta\zeta^l \simeq 1$ for large l and small diffusion, i.e. large width. This is further supported by the results of our experiments in Fig. 1a and Fig. 1b. A direct consequence of $\delta\zeta^l \simeq 1$ is that the ratio $\nu_{1,\text{c}}(\mathbf{y}^{l,+})\nu_{1,\text{c}}(\mathbf{y}^{l,-}) \,/\, \mu_2(\mathbf{x}^{l-1})$ in Eq. (8) is constrained to small values, which leaves two possibilities for the *signal distribution* that we illustrate qualitatively:

(i) If $\max\left(\nu_{1,\text{c}}(\mathbf{y}^{l,+}), \nu_{1,\text{c}}(\mathbf{y}^{l,-})\right)^2 \,/\, \mu_2(\mathbf{x}^{l-1}) \to 0$, then $\nu_{1,c}(|\mathbf{y}|)^2 \,/\, \mu_2(\mathbf{x}^{l-1}) \to 0$ and first-order standardized moments become ill-defined.

(ii) If $\min\left(\nu_{1,c}(\mathbf{y}^{l,+}), \nu_{1,c}(\mathbf{y}^{l,-})\right)^2 / \mu_2(\mathbf{x}^{l-1}) \to 0$, then $\mathbf{f}_m(\mathbf{y}^l, \boldsymbol{\alpha})$ becomes concentrated on the semi-line generated by its average vector $(\nu_{1,c}(\mathbf{y}^l))_{1\leq c\leq N_l}$ (proof in Appendix F.3). In this case, the same pattern of activation is seen with probability one with respect to input data and the neural network becomes *linear*.

The situation (ii) is clearly visible in Fig. 1c. Our analysis provides a novel insight in this previously observed phenomenon of coactivation (Balduzzi et al., 2017). Note that the distributional pathology is severe since a 1-dimensional distribution at layer l implies a 1-dimensional input when taking the perspective of layers $l' > l$.

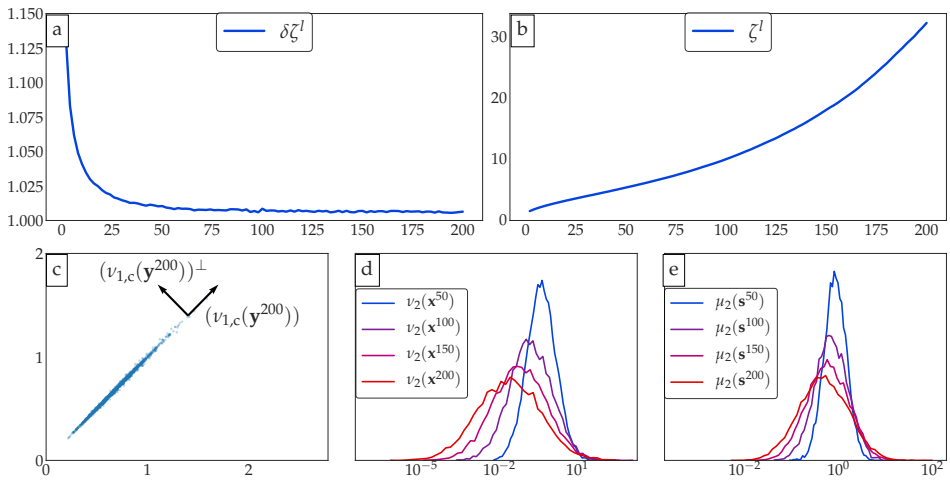

Figure 1: Illustration of the distributional pathologies of vanilla networks with $N_l = 128$ and $L = 200$ layers. *(a)* Geometric increments of the normalized sensitivity $\delta\zeta^l$ with rapid evolution towards $\delta\zeta^l \simeq 1$. *(b)* The normalized sensitivity ζ^l has sub-exponential evolution since it is limited by neural network pseudo-linearity. *(c)* 2-dimensional cuts of preactivation feature maps $\mathbf{f}_m(\mathbf{y}^{200}, \boldsymbol{\alpha})$ using the direction of average vector $(\nu_{1,c}(\mathbf{y}^{200}))$ and a random orthogonal direction $(\nu_{1,c}(\mathbf{y}^{200}))^\perp$. *(d-e)* Evolution of the distributions of $\mu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ with depth showing clear lognormality.

6 Batch-Normalized Feedforward Nets

Next we incorporate batch normalization (Ioffe & Szegedy, 2015) which has the effect of subtracting $\nu_{1,c}(\mathbf{y}^l)$ and normalizing by $\mu_{2,c}(\mathbf{y}^l)^{1/2}$ for each channel c in $\mathbf{y}^l$. The equations of propagation are given by

$$\mathbf{y}^l = \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l, \quad \mathbf{z}^l = BN(\mathbf{y}^l), \qquad \mathbf{x}^l = \phi(\mathbf{z}^l), \tag{9}$$

$$\mathbf{t}^l = \boldsymbol{\omega}^l * \mathbf{s}^{l-1}, \qquad \mathbf{u}^l = BN'(\mathbf{y}^l) \odot \mathbf{t}^l, \quad \mathbf{s}^l = \phi'(\mathbf{z}^l) \odot \mathbf{u}^l, \tag{10}$$

where BN denotes batch normalization and where we introduced the tensors $\mathbf{z}^l$, $\mathbf{t}^l$ and $\mathbf{u}^l$. Note that Eq. (9) and Eq. (10) explicitly formulate a finer-grained subdivision of three different steps between layers $l-1$ and l in the simultaneous propagation of $(\mathbf{x}^l, \mathbf{s}^l)$.

Theorem 3. Normalized Sensitivity increments of batch-normalized feedforward nets. (proof in Appendix G.1) *The dominating term in the evolution of ζ^l can be decomposed as the sum of a term $\overline{m}_{BN}[\zeta^l]$ due to batch normalization and a term $\overline{m}_\phi[\zeta^l]$ due to the nonlinearity ϕ:*

$$\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right) = \left(\frac{\mu_2(\mathbf{s}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2} \mathbb{E}_{c,\theta^l}\left[\frac{\mu_{2,c}(\mathbf{t}^l)}{\mu_{2,c}(\mathbf{y}^l)}\right]^{1/2}, \tag{11}$$

$$\exp\left(\overline{m}_\phi\left[\zeta^l\right]\right) = \left(1 - 2\mathbb{E}_{c,\theta^l}\left[\nu_{1,c}(\mathbf{z}^{l,+})\nu_{1,c}(\mathbf{z}^{l,-})\right]\right)^{-1/2}, \tag{12}$$

$$\delta\zeta^l \simeq \exp\left(\overline{m}_{BN/FF}\left[\zeta^l\right]\right) = \exp\left(\overline{m}_{BN}\left[\zeta^l\right] + \overline{m}_\phi\left[\zeta^l\right]\right). \tag{13}$$

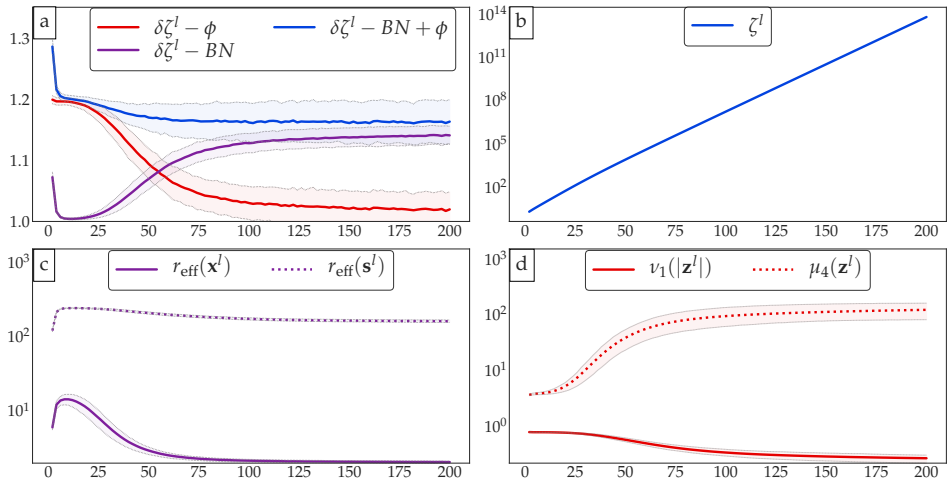

Figure 2: Illustration of the distributional pathology of batch-normalized feedforward nets with $N_l = 384$ and $L = 200$ layers. *(a)* Geometric increments $\delta\zeta^l$ and their decomposition as the product of two terms: a batch normalization term corresponding to the evolution from $(\mathbf{x}^{l-1}, \mathbf{s}^{l-1})$ to $(\mathbf{z}^l, \mathbf{u}^l)$ and a nonlinearity term corresponding to the evolution from $(\mathbf{z}^l, \mathbf{u}^l)$ to $(\mathbf{x}^l, \mathbf{s}^l)$. *(b)* The normalized sensitivity ζ^l has exploding behaviour. *(c)* Effective ranks of signal and sensitivity confirm that sensitivity is much better conditioned than signal. There is a clear inverse correlation between $r_{\text{eff}}(\mathbf{x}^l)$ and the batch normalization term in $\delta\zeta^l$. *(d)* $\mathbf{z}^l$ becomes ill-behaved as $\nu_1(|\mathbf{z}^l|)$ vanishes and $\mu_4(\mathbf{z}^l)$ explodes. This explains the decay of the nonlinearity term in $\delta\zeta^l$.

Effect of batch normalization. The term of Eq. (11) corresponds to the evolution of ζ^l from $(\mathbf{x}^{l-1}, \mathbf{s}^{l-1})$ at layer $l-1$ to $(\mathbf{z}^l, \mathbf{u}^l)$ just after BN. To understand this term qualitatively, the pre-activation tensor $\mathbf{y}^l$ can be seen as N_l random projections of $\mathbf{x}^{l-1}$, and batch normalization can be seen as an alteration of the magnitude for each projection. Given that batch normalization uses $\mu_{2,\text{c}}(\mathbf{y}^l)^{1/2}$ as normalization factor, directions of high signal variance are dampened while directions of low signal variance are amplified. This *preferential exploration* of low signal directions naturally deteriorates the signal-to-noise ratio and amplifies ζ^l due to the factor of noise equivalence.

Now let us look directly at the quantity inside the expectation in Eq. (11). By spherical symmetry under standard initialization, geometric increments from $\mathbf{x}^{l-1}$ to $\mathbf{y}^l$ for the signal and $\mathbf{s}^{l-1}$ to $\mathbf{t}^l$ for the sensitivity have the same expectation $\mathbb{E}_{\text{c},\theta^l}[\mu_{2,\text{c}}(\mathbf{t}^l)] \,/\, \mu_{2,\text{c}}(\mathbf{s}^{l-1}) = \mathbb{E}_{\text{c},\theta^l}[\mu_{2,\text{c}}(\mathbf{y}^l)] \,/\, \mu_{2,\text{c}}(\mathbf{x}^{l-1})$. On the other hand, the fluctuation of these geometric increments depends on the fluctuation of the signal and sensitivity in the N_l random projections, i.e. on whether directions of signal and sensitivity variances are rare in the ambient space. To measure this effect of *conditioning*, we adopt the metric of effective rank r_{eff} from Vershynin (2012):

$$r_{\text{eff}}(\mathbf{x}^l) \equiv \frac{\text{Tr}\, \boldsymbol{C}\left[\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})\right]}{||\boldsymbol{C}\left[\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})\right]||}, \qquad r_{\text{eff}}(\mathbf{s}^l) \equiv \frac{\text{Tr}\, \boldsymbol{C}\left[\mathbf{f}_{\text{m}}(\mathbf{s}^l, \boldsymbol{\alpha})\right]}{||\boldsymbol{C}\left[\mathbf{f}_{\text{m}}(\mathbf{s}^l, \boldsymbol{\alpha})\right]||}, \tag{14}$$

whith $\boldsymbol{C}\left[\mathbf{f}_{\text{m}}(\mathbf{x}^l, \boldsymbol{\alpha})\right]$, $\boldsymbol{C}\left[(\mathbf{f}_{\text{m}}(\mathbf{s}^l, \boldsymbol{\alpha})\right]$ the covariance matrices of signal and sensitivity feature map vectors and $||\cdot||$ the spectral norm. We further extend this definition to any random tensor. If we assume that $\mathbf{s}^l$ has very good conditioning with $r_{\text{eff}}(\mathbf{s}^{l-1}) \simeq N_{l-1}$, then $\mu_{2,\text{c}}(\mathbf{t}^l)$ has small relative deviation to its expectation $\mathbb{E}_{\text{c},\theta^l}[\mu_{2,\text{c}}(\mathbf{t}^l)]$ and this term can be treated as a constant. In turn, this implies by convexity of $x \to 1/x$ that $\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right) \gtrsim 1$. The worse the conditioning of $\mathbf{x}^{l-1}$, i.e. the smaller $r_{\text{eff}}(\mathbf{x}^{l-1})$, the larger the variance of $\mu_{2,\text{c}}(\mathbf{y}^l) \,/\, \mathbb{E}_{\text{c},\theta^l}[\mu_{2,\text{c}}(\mathbf{y}^l)]$ at the denominator and the impact of the convexity. Thus the smaller $r_{\text{eff}}(\mathbf{x}^{l-1})$ and the larger $\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right)$. This argument is strictly valid for the first step of the propagation where the sensitivity has perfect conditioning, which results in $\exp\left(\overline{m}_{BN}\left[\zeta^1\right]\right) \geq 1$ (proof in Appendix G.2). In Fig. 2c we confirm experimentally that $r_{\text{eff}}(\mathbf{s}^l) \simeq N_l \gg r_{\text{eff}}(\mathbf{x}^l)$. Together with Fig. 2a we also confirm that $r_{\text{eff}}(\mathbf{x}^l)$ is highly predictive of the batch normalization effect on $\delta\zeta^l$.

Effect of the nonlinearity ϕ. The term of Eq. (12) corresponds to the evolution of ζ^l from $(\mathbf{z}^l, \mathbf{u}^l)$ after BN to $(\mathbf{x}^l, \mathbf{s}^l)$ after ϕ. The quantity inside the expectation can also be expressed as $\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-}) \,/\, \mu_{2,\mathrm{c}}(\mathbf{z}^l)$ since $\mu_{2,\mathrm{c}}(\mathbf{z}^l) = 1$ after batch normalization. We then find a very similar expression as Eq. (8) for vanilla networks. The difference is that first-order moments are now normalized using $\mu_{2,\mathrm{c}}(\mathbf{z}^l)$ instead of $\mu_2(\mathbf{x}^{l-1})$. This implies that each random projection in $\mathbf{y}^l$ is given the same importance in the sense of similar contribution to $\delta\zeta^l$. On the contrary, Eq. (8) for vanilla networks gives more importance to random directions with high signal since $\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) \,/\, \mu_2(\mathbf{x}^{l-1})$ is small for low signal directions. Note that $\nu_{1,\mathrm{c}}(|\mathbf{z}^l|) = \nu_{1,\mathrm{c}}(\mathbf{z}^{l,+}) + \nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})$ implies $2\,\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-}) \leq \nu_{1,\mathrm{c}}(|\mathbf{z}^l|)^2$ when taking the square. The term $\overline{m}_\phi[\zeta^l]$ is thus limited by $\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(|\mathbf{z}^l|)^2]$ as shown by the joint examination of Fig. 2a and Fig. 2d.

7 Batch-Normalized Resnets

We finish our exploration of DNN architectures with the incorporation of skip connections. We now suppose that the width is constant, i.e. $N_l = N$,[4] and following He et al. (2016) we adopt the perspective of *pre-activation units*. The propagation inside residual units is given by

$$\mathbf{y}^{l,h} = \omega^{l,h} * \phi\Big(BN\big(\mathbf{y}^{l,h-1}\big)\Big) + \beta^{l,h}, \tag{15}$$

$$\mathbf{t}^{l,h} = \omega^{l,h} * \Big(\mathbf{t}^{l,h-1} \odot BN'\big(\mathbf{y}^{l,h-1}\big) \odot \phi'\Big(BN\big(\mathbf{y}^{l,h-1}\big)\Big)\Big), \tag{16}$$

where Eq. (15) and Eq. (16) hold for $1 \leq h \leq H$, with H the number of layers in each residual unit. Denoting $(\mathbf{y}^{l,h}, \mathbf{t}^{l,h}) = \Phi^{l,h}(\mathbf{y}^{l,h-1}, \mathbf{t}^{l,h-1})$, the propagation between successive residual units is given by

$$(\mathbf{y}^l, \mathbf{t}^l) = (\mathbf{y}^{l-1}, \mathbf{t}^{l-1}) + (\mathbf{y}^{l,H}, \mathbf{t}^{l,H}) = (\mathbf{y}^{l-1}, \mathbf{t}^{l-1}) + \Phi^{l,H} \ldots \Phi^{l,1}(\mathbf{y}^{l-1}, \mathbf{t}^{l-1}). \tag{17}$$

For consistency reasons, we rename the inputs of the propagation as $\mathbf{y}^0 \equiv \mathbf{y}$, $\mathbf{t}^0 \equiv \mathbf{t}$. We further adopt the convention that $\mathbf{y}^{0,H} \equiv \mathbf{y}^0$, $\mathbf{t}^{0,H} \equiv \mathbf{t}^0$ such that Eq. (17) can be rewritten as

$$(\mathbf{y}^l, \mathbf{t}^l) = \sum_{k=0}^{l} (\mathbf{y}^{k,H}, \mathbf{t}^{k,H}). \tag{18}$$

Theorem 4. Normalized Sensitivity increments of batch-normalized resnets. (proof in Appendix H.3) *Suppose that for all depth l we can bound the effective ranks $r_{\min} \lesssim r_{\text{eff}}(\mathbf{y}^l), r_{\text{eff}}(\mathbf{y}^{l,H}), r_{\text{eff}}(\mathbf{t}^l), r_{\text{eff}}(\mathbf{t}^{l,H})$, the second-order central moment $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and the feedforward increments inside residual units $\delta_{\min} \lesssim \delta\zeta^{l,h} \lesssim \delta_{\max}$. Denote $\rho_{\min} = \big((\delta_{\min})^{2H}\mu_{2,\min} - \mu_{2,\max}\big) \,/\, \mu_{2,\max}$ and $\rho_{\max} = \big((\delta_{\max})^{2H}\mu_{2,\max} - \mu_{2,\min}\big) \,/\, \mu_{2,\min}$, and further consider $\tau_{\min}$, $\tau_{\max}$ such that $\tau_{\min} < \rho_{\min} \,/\, 2$ and $\tau_{\max} > \rho_{\max} \,/\, 2$.[5] Then:*

$$\forall l \ll N r_{\min} : \quad \Big(1 + \frac{\rho_{\min}}{l+1}\Big)^{1/2} \lesssim \delta\zeta^l \lesssim \Big(1 + \frac{\rho_{\max}}{l+1}\Big)^{1/2}, \tag{19}$$

$$\forall l \gg 1 : \quad \frac{1}{2}\rho_{\min}\log l \lesssim \log\zeta^l \lesssim \frac{1}{2}\rho_{\max}\log l, \tag{20}$$

$$\forall l \gg 1 : \quad l^{\tau_{\min}} \lesssim \zeta^l \lesssim l^{\tau_{\max}}. \tag{21}$$

Discussion. The evolution in Eq. (19) remains exponential inside residual units since $\rho_{\min}$ and $\rho_{\max}$ have an exponential dependence in H. However it is slowed down by the factor $1 \,/\, (l+1)$ between successive residual units. This comes from the *dilution* of the residual path $(\mathbf{y}^{l,H}, \mathbf{t}^{l,H})$ in the skip connection path $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ with ratio of signal variances $\mu_2(\mathbf{y}^{l,H}) \,/\, \mu_2(\mathbf{y}^{l-1})$ scaling as $1 \,/\, l$. If we remove the dilution effect in Eq. (19) by replacing $l+1$ by 1 and if we set $\mu_{2,\min} = \mu_{2,\max}$,

[4] Again this assumption is only made for simplicity of the analysis. In practice, it holds at least approximately since N_l is only modified by very few units.

[5] Note that Theorem 4 has very mild assumptions. The assumption on effective ranks is not required in Eq. (20) and Eq. (21). Furthermore the assumption on $\mu_2(\mathbf{y}^{l,H})$ is very reasonable since batch normalization controls signal variance at the beginning of layer H.

then we recover the feedforward evolution with $(\delta_{\text{min}})^H \lesssim \zeta^l \lesssim (\delta_{\text{max}})^H$. The dilution is clearly visible as a side effect of the layer aggregation in Eq. (18): each residual unit l adds a new term of increased sensitivity, but its relative contribution to the aggregation becomes smaller and smaller with l, so it gets harder and harder for the model to grow ζ^l.

Since $\delta\zeta^l$ becomes closer and closer to 1, its fluctuation eventually becomes dominant relatively to its expected deviation to 1. This explains why Eq. (19) only holds for small l. It continues however to hold statistically so that the bounds on $\log \zeta^l = \sum_k \log \delta\zeta^k$ in Eq. (20) correspond to the integration of the bounds in Eq. (19). A direct consequence of the dilution is thus the power-law evolution of ζ^l in Eq. (21) instead of the exponential evolution for feedforward nets. Equivalently, when Eq. (21) is written as $\exp\left(\tau_{\text{min}} \log l\right) \lesssim \zeta^l \lesssim \exp\left(\tau_{\text{max}} \log l\right)$, the evolution of ζ^l for resnets is the same as the evolution of $\zeta^{\log l}$ for feedforward nets. In words, the evolution with depth of resnets is the *logarithmic* version of the evolution with depth of feedforward nets. Up to some factor, an evolution from 100, to 1 000, and to 10 000 layers for resnets is equivalent to an evolution from 20, to 30, and to 40 layers for feedforward nets. Despite differences in the underlying assumptions, our results are reminiscent of the results in Philipp et al. (2017) on the role of the dilution to alleviate the exploding gradient problem, as well as the results in Yang & Schoenholz (2017) on the power-law evolution of mean-field resnet gradients.

A shown in Fig. 3, the slow power-law evolution ensures that all statistical quantities remain well-behaved. The exponent in the power-law fit in Fig. 3d is set to $\tau = \rho\,/\,2 = (\delta_{\text{av}}^{2H} - 1)\,/\,2$, with δ_{av} the feedforward increment averaged over the whole evolution. This shows that the bound in Eq. (21) is tight.

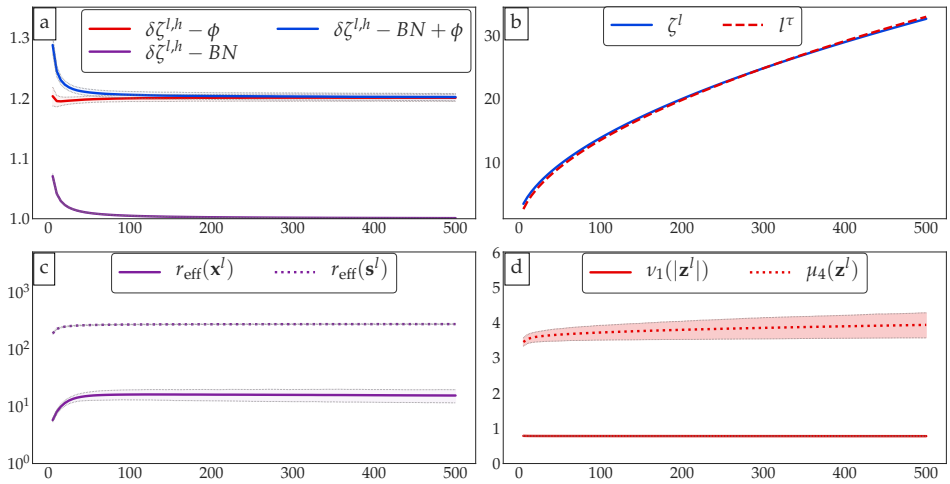

Figure 3: Illustration of the well-behaved evolution of batch-normalized resnets with $N_l = 384$ and $L = 500$ residual units of $H = 2$ layers. *(a)* Decomposing $\delta\zeta^{l,h}$ as the product of two terms: a batch normalization term and a nonlinearity term. *(b)* The normalized sensitivity ζ^l has power law evolution. *(c)* Effective ranks of signal and sensitivity indicate that many directions of signal variance are preserved. *(d)* Moments $\nu_1(|\mathbf{z}^l|)$ and $\mu_4(\mathbf{z}^l)$ indicate that $\mathbf{z}^l$ has nearly Gaussian distribution.

8 Conclusion

This paper introduced a novel approach for the statistical characterization of deep neural networks and their sensitivity. Only very mild assumptions were required and most ingredients of modern DNNs were incorporated. The main scope restriction comes from our focus on the rectifier activation function. We expect however qualitatively similar results to hold for other activation functions.

Below is a summary of our key results:

– For vanilla networks, He et al. (2015) only stabilizes second-order moments of activation and sensitivity in expectation. Depth propagation still induces an additive random walk with small negative drift in log-space. This results in slowly vanishing activations and gradients and the

inevitable convergence to a distributional pathology where the neural network becomes linear and its signal shrunk to a single dimension.

– For batch-normalized feedforward nets, the exponential growth of sensitivity has two origins: on the one hand batch normalization which upweights low-signal pre-activation directions, on the other hand the nonlinear activation function ϕ.

– Finally for resnets the sensitivity only grows as a power-law. Equivalently the evolution with depth of resnets is the logarithmic version of the evolution with depth of feedforward nets. The underlying phenomenon is the dilution of the residual path in the skip connection path with ratio of signal variances decaying as $1 \,/\, k$. This ingenious mechanism is responsible for breaking the circle of depth multiplicativity which causes distributional pathology for vanilla networks and batch-normalized feedforward nets.

We hope that our methodology will open new perspectives in the statistical understanding of deep neural network architectures. We believe that it could also provide novel insights regarding model trainability and generalization.

REFERENCES

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. The shattered gradients problem: If resnets are the answer, then what is the question? In *International Conference on Machine Learning*, 2017.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *ArXiv e-prints*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1026–1034. IEEE Computer Society, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, pp. 630–645. Springer International Publishing, 2016.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 448–456, 2015.

L. Lu, Y. Su, and G. E. Karniadakis. Collapse of Deep and Narrow Neural Nets. *ArXiv e-prints*, 2018.

Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring generalization in deep learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5947–5956. 2017.

R. Novak, Y. Bahri, D. A. Abolafia, J. Pennington, and J. Sohl-Dickstein. Sensitivity and Generalization in Neural Networks: an Empirical Study. In *International Conference on Learning Representations*, 2018.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. *CoRR*, abs/1711.04735, 2017.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. In *AISTATS*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1924–1932. PMLR, 2018.

G. Philipp and J. G. Carbonell. The Nonlinearity Coefficient - Predicting Overfitting in Deep Neural Networks. *ArXiv e-prints*, 2018.

George Philipp, Dawn Song, and Jaime G. Carbonell. Gradients explode - deep networks are shallow - resnet explained. ArXiv e-prints, 2017.

Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 3360–3368. Curran Associates, Inc., 2016.

Maithra Raghu, Ben Poole, Jon M. Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854. PMLR, 2017.

Samuel S. Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *CoRR*, abs/1611.01232, 2016.

Jure Sokolic, Raja Giryes, Guillermo Sapiro, and Miguel R. D. Rodrigues. Robust large margin deep neural networks. *IEEE Trans. Signal Processing*, 65(16):4265–4280, 2017.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012.

L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. *ArXiv e-prints*, 2018.

Ge Yang and Samuel Schoenholz. Mean field residual networks: On the edge of chaos. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 7103–7114. Curran Associates, Inc., 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv e-prints*, 2016.

A Details of the experiments in Fig. 1, Fig. 2 and Fig. 3

All three experiments of Fig. 1, Fig. 2 and Fig. 3 were made on CIFAR10 with a random initial convolution of stride 2 reducing the spatial dimension to $n = 16$. In each case we considered the convolutional extent $K_l = 3$ and periodic boundary conditions. A few experiments with approximately statistics-preserving boundary conditions such as *symmetric mirroring* indicated qualitatively equivalent behaviour.

Fig. 1 was obtained by considering width $N_l = 128$ and total depth $L = 200$. For 10 000 random initializations we randomly sampled 1 024 images and computed the evolution with depth of $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$. The distributions of $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ were estimated with the empirical distributions on the 10 000 computations, which are shown in Fig. 1d and Fig. 1e. As for Fig. 1c, it was obtained with 2-dimensional cuts of preactivation feature maps $\mathbf{f}_{\mathrm{m}}(\mathbf{y}^{200}, \boldsymbol{\alpha})$ for 1 randomly sampled point every 5 computations. Each time the cut was performed using the direction of average vector $(\nu_{1,\mathrm{c}}(\mathbf{y}^{200}))$ and a random orthogonal direction $(\nu_{1,\mathrm{c}}(\mathbf{y}^{200}))^{\perp}$.

Fig. 2 was obtained by considering 1 000 batches of 64 randomly sampled images. Due to less demanding computations, we increased the width to a more realistic value $N_l = 384$. For simplicity, we always considered batch normalization in train mode. Solid curves correspond to the expectation over the 1 000 batches and Fig. 2a, Fig. 2c, Fig. 2d additionally show 1σ intervals.

Finally Fig. 3 was obtained by considering again 1 000 batches of 64 randomly sampled images with $N_l = 384$ and $L = 500$ residual units of $H = 2$ layers. Geometric increments $\delta\zeta^{l,h}$ were computed in the feedforward propagation from $(\mathbf{y}^{l,h-1}, \mathbf{t}^{l,h-1})$ to $(\mathbf{y}^{l,h}, \mathbf{t}^{l,h})$ at layer $h = 2$ in residual unit l. Solid curves correspond again to the expectation over the 1 000 batches and Fig. 3a, Fig. 3c, Fig. 3d additionally show 1σ intervals.

We plan to release Jupyter notebooks to enable replication of our results upon publication.

B COMPLEMENTARY DEFINITIONS AND NOTATIONS

Receptive field mapping. Here we temporarily need to handle the mechanics of convolution. Let us consider the convolution at layer l of an input $\mathbf{v} \in \mathbb{R}^{n\times\cdots\times n\times N_{l-1}}$ from layer $l-1$. The output feature map of the convolution $(w^l * \mathbf{v})_{\boldsymbol{\alpha},:}$ at position $\boldsymbol{\alpha} \in \{1,\dots,n\}^d$ is obtained by the application of the convolution kernel w^l over a local input region of size $(K_l^d N_{l-1})$, with K_l^d the spatial extent and N_{l-1} the extent in the channel dimension. The local input region is called the *receptive field* of $(w^l * \mathbf{v})$ at spatial position $\boldsymbol{\alpha}$.

The *receptive field mapping* associates an input $\mathbf{v}$ from layer $l-1$ to $RF(\mathbf{v})$. $RF(\mathbf{v})$ is the tensor of $\mathbb{R}^{n\times\cdots\times n\times K_l^d N_{l-1}}$ such that $RF(\mathbf{v})_{\boldsymbol{\alpha},:}$ is the reshaped vectorial form of the receptive field of $(w^l * \mathbf{v})$ at spatial position $\boldsymbol{\alpha}$. We denote $R_l = K_l^d N_{l-1}$ the dimensionality of $RF(\mathbf{v})_{\boldsymbol{\alpha},:}$ and $\mathcal{I}_{\mathrm{c}}^l$ the set of indices in $RF(\mathbf{v})_{\boldsymbol{\alpha},:}$ corresponding to elements in channel c in the input $\mathbf{v}$. Strictly speaking, RF depend on l, but this is implied by the argument so we write RF for simplicity.

Receptive field vectors. The *receptive field vector* $\mathbf{r}_{\mathrm{f}}$ and *centered receptive field vector* $\hat{\mathbf{r}}_{\mathrm{f}}$ associated with an input $\mathbf{v}$ from layer l are random vectors which depend on $\mathbf{v}$, $\boldsymbol{\alpha}$ such that

$$\mathbf{r}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha}) \equiv RF(\mathbf{v})_{\boldsymbol{\alpha},:} \quad \text{and} \quad \hat{\mathbf{r}}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha}) \equiv \mathbf{r}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha}) - \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha})],$$

where we kept the same denotation for the variable in the expectation, as a slight abuse of notation. Again $\mathbf{r}_{\mathrm{f}}$ and $\hat{\mathbf{r}}_{\mathrm{f}}$ are strictly speaking dependent on l, but this is implied by the argument.

Statistics-preserving property. RF is *statistics-preserving* with respect to $\mathbf{v}$ if for any channel c and any index $i_{\mathrm{c}} \in \mathcal{I}_{\mathrm{c}}^l$, the random variables $RF(\mathbf{v})_{\boldsymbol{\alpha},i_{\mathrm{c}}}$ and $\mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ which depend on $\mathbf{v}$, $\boldsymbol{\alpha}$ have the same distribution $RF(\mathbf{v})_{\boldsymbol{\alpha},i_{\mathrm{c}}} \overset{\mathbf{v},\boldsymbol{\alpha}}{\sim} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$.

Equation of Propagation. Using the definition of RF, the affine transformation from the receptive field $RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ to the feature map in the next layer $\mathbf{y}^l_{\boldsymbol{\alpha},:}$ is written as

$$\mathbf{y}^l_{\boldsymbol{\alpha},:} = \boldsymbol{W}^l RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} + \mathbf{b}^l, \tag{22}$$

where $\boldsymbol{W}^l \in \mathbb{R}^{N_l \times R_l}$ is the suitably reshaped matricial form of $\boldsymbol{\omega}^l$. To lighten notation, we write $\mathbf{y}^l = \boldsymbol{W}^l RF(\mathbf{x}^{l-1}) + \mathbf{b}^l$ as a short for the affine transformation of Eq. (22) occuring at all spatial positions $\boldsymbol{\alpha}$. We have the following equivalence between the notations with receptive field and with convolutions:

$$\boldsymbol{W}^l RF(\mathbf{x}^{l-1}) + \mathbf{b}^l = \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l.$$

For vanilla networks, the simultaneous propagation of $\mathbf{x}^l$ and $\mathbf{s}^l$ is then written as

$$\mathbf{y}^l = \boldsymbol{W}^l RF(\mathbf{x}^{l-1}) + \mathbf{b}^l, \quad \mathbf{x}^l = \phi(\mathbf{y}^l), \tag{23}$$

$$\mathbf{t}^l = \boldsymbol{W}^l RF(\mathbf{s}^{l-1}), \qquad \mathbf{s}^l = \phi'(\mathbf{y}^l) \odot \mathbf{t}^l. \tag{24}$$

For batch-normalized feedforward nets, the simultaneous propagation of $\mathbf{x}^l$ and $\mathbf{s}^l$ is written as

$$\begin{aligned} \mathbf{y}^l &= \boldsymbol{W}^l RF(\mathbf{x}^{l-1}) + \mathbf{b}^l, & \mathbf{z}^l &= BN(\mathbf{y}^l), & \mathbf{x}^l &= \phi(\mathbf{z}^l), \\ \mathbf{t}^l &= \boldsymbol{W}^l RF(\mathbf{s}^{l-1}), & \mathbf{u}^l &= BN'(\mathbf{y}^l) \odot \mathbf{t}^l, & \mathbf{s}^l &= \phi'(\mathbf{z}^l) \odot \mathbf{u}^l. \end{aligned}$$

Covariance and Gramian. The *Gramian operator* $\boldsymbol{G}$ and *covariance operator* $\boldsymbol{C}$ associated with a random vector $\mathbf{v} \in \mathbb{R}^N$ are defined as

$$\boldsymbol{G}[\mathbf{v}] \equiv \mathbb{E}_{\mathbf{v}}\left[\mathbf{v}\mathbf{v}^T\right] \quad \text{and} \quad \boldsymbol{C}[\mathbf{v}] \equiv \mathbb{E}_{\mathbf{v}}\left[\mathbf{v}\mathbf{v}^T\right] - \mathbb{E}_{\mathbf{v}}\left[\mathbf{v}\right]\mathbb{E}_{\mathbf{v}}\left[\mathbf{v}\right]^T.$$

Note that the gramian and covariance of feature map and receptive field vectors are related by $\boldsymbol{G}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v},\boldsymbol{\alpha})] = \boldsymbol{C}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v},\boldsymbol{\alpha})] = \boldsymbol{C}[\mathbf{f}_{\mathrm{m}}(\mathbf{v},\boldsymbol{\alpha})]$ and $\boldsymbol{G}[\hat{\mathbf{r}}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha})] = \boldsymbol{C}[\hat{\mathbf{r}}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha})] = \boldsymbol{C}[\mathbf{r}_{\mathrm{f}}(\mathbf{v},\boldsymbol{\alpha})]$.

Symmetric propagation for vanilla networks. We define additional tensors obtained by *symmetric propagation* at each layer l. In the case of vanilla networks they are given by:

$$\tilde{\mathbf{y}}^l = -\boldsymbol{W}^l RF(\mathbf{x}^{l-1}) - \mathbf{b}^l, \quad \tilde{\mathbf{x}}^l = \phi(\tilde{\mathbf{y}}^l),$$
$$\tilde{\mathbf{t}}^l = -\boldsymbol{W}^l RF(\mathbf{s}^{l-1}), \qquad \tilde{\mathbf{s}}^l = \phi'(\tilde{\mathbf{y}}^l) \odot \tilde{\mathbf{t}}^l.$$

By spherical symmetry, tensor moments have the *same distribution with respect to θ^l for both propagations*. Furthermore $\forall \boldsymbol{\alpha}, \mathrm{c}$, $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} + \tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = \left|\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}\right|$ and $\left(\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}\right)^2 + \left(\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}}\right)^2 = \left(\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}\right)^2$ since $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$. We deduce

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{y}^l), \tag{25}$$

Now consider the second-order moments of the sensitivity tensor and suppose first that $\mathbf{x}, \mathbf{s}, \boldsymbol{\alpha}, \mathrm{c}, \Theta^{l-1}$ are fixed. We have the following identity:

$$(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{t}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\phi'(\tilde{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2[\phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + \phi'(\tilde{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2].$$

There are two possible cases depending on $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} = \boldsymbol{W}^l_{\mathrm{c},:} RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} + \mathbf{b}^l$:

- If $||RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}||_2^2 \neq 0$, then under standard initialization, $\mathbb{P}_{\theta^l}\left[\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} \neq 0\right] = 1$, and thus $\mathbb{P}_{\theta^l}\left[\phi'(\mathbf{y}^l)^2 + \phi'(\tilde{\mathbf{y}}^l)^2 = 1\right] = 1$ and $\mathbb{P}_{\theta^l}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = 1$.
- If $||RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}||_2^2 = 0$, then the element-wise tensor multiplication of Eq. (24) at layer $l-1$ implies $||RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:}||_2^2 = 0$, and thus $\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$, $\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$, $\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$.

In all cases, $\mathbb{P}_{\theta^l}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = 1$. For given c, we now consider $\mathbf{x}, \mathbf{s}, \boldsymbol{\alpha}$ as variables again and we get by Fubini's theorem:

$$\mathbb{E}_{\theta^l}\mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\mathbf{1}_{\left\{(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right\}}\right] = 1,$$
$$\mathbb{E}_{\theta^l}\mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\mathbf{1}_{\left\{(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \neq (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right\}}\right] = 0, \tag{26}$$
$$\mathbb{P}_{\theta^l}\left[\mathbb{P}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \neq (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = 0\right] = 1. \tag{27}$$

where Eq. (27) is obtained by contradiction, since $\mathbb{P}_{\theta^l}\left[\mathbb{P}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \neq (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] > 0\right] > 0$ would imply that Eq. (26) does not hold. We can relate the moments of $\mathbf{s}^l$, $\tilde{\mathbf{s}}^l$ and $\mathbf{t}^l$ by

$$\begin{aligned}
\left(\nu_{2,\mathrm{c}}(\mathbf{s}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) - \nu_{2,\mathrm{c}}(\mathbf{t}^l)\right)^2 &= \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 - (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right]^2 \\
&= \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\left((\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 - (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right)\mathbf{1}_{\left\{(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \neq (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right\}}\right]^2 \\
&\leq \nu_{2,\mathrm{c}}\left((\mathbf{s}^l)^2 + (\tilde{\mathbf{s}}^l)^2 - (\mathbf{t}^l)^2\right)\mathbb{P}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \neq (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right],
\end{aligned}$$

where we used Cauchy-Schwarz inequality and the implicit assumption that the moment quantity is well-defined. Combined with Eq. (27), we then get

$$\mathbb{P}_{\theta^l}\left[\nu_{2,\mathrm{c}}(\mathbf{s}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \nu_{2,\mathrm{c}}(\mathbf{t}^l)\right] = 1.$$

Since $\mathbf{s}^l$, $\tilde{\mathbf{s}}^l$ and $\mathbf{t}^l$ are centered, $\nu_{2,\mathrm{c}}(\mathbf{s}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l)$ and $\nu_{2,\mathrm{c}}(\mathbf{t}^l) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)$. So we finally get for the event intersection $\left\{\forall \mathrm{c} \ : \ \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\} = \bigcap_{\mathrm{c}=1}^{N_l}\left\{\mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\}$:

$$\mathbb{P}_{\theta^l}\left[\left\{\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\}\right] = 1. \tag{28}$$

Symmetric propagation for batch-normalized feedforward nets. For batch-normalized feedforward nets, the symmetric propagation at each layer l is given by

$$\tilde{\mathbf{y}}^l = -\boldsymbol{W}^l RF(\mathbf{x}^{l-1}) - \mathbf{b}^l, \quad \tilde{\mathbf{z}}^l = BN(\tilde{\mathbf{y}}^l), \quad \tilde{\mathbf{x}}^l = \phi(\tilde{\mathbf{z}}^l), \tag{29}$$
$$\tilde{\mathbf{t}}^l = -\boldsymbol{W}^l RF(\mathbf{s}^{l-1}), \quad \tilde{\mathbf{u}}^l = BN'(\tilde{\mathbf{y}}^l) \odot \tilde{\mathbf{t}}^l, \quad \tilde{\mathbf{s}}^l = \phi'(\tilde{\mathbf{z}}^l) \odot \tilde{\mathbf{u}}^l. \tag{30}$$

BN in Eq. (29) and Eq. (30) use the statistics of $\tilde{\mathbf{y}}^l$, so that tensor moments have the *same distribution with respect to θ^l for both propagations*. We then simply have

$$\tilde{\mathbf{z}}^l = -\mathbf{z}^l, \qquad \tilde{\mathbf{u}}^l = -\mathbf{u}^l. \tag{31}$$

The exact same analysis as before gives:

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{z}^l), \tag{32}$$

$$\mathbb{P}_{\theta^l}\Big[\big\{\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{u}^l)\big\}\Big] = 1. \tag{33}$$

C Statistics-Preserving Property

C.1 Case of periodic boundary conditions and constant spatial extent n

Lemma 1. *If convolutions have periodic boundary conditions and the global spatial extent n is constant, then RF is statistics-preserving with respect to any input $\mathbf{v}$.*

Proof. Fix a channel c in $\mathbf{v}$, an index $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$, and consider the tensors $\mathbf{v}_{:,\mathrm{c}}$ and $RF(\mathbf{v})_{:,i_\mathrm{c}} \in \mathbb{R}^{n\times\cdots\times n}$. The index i_c corresponds to a given convolution kernel position $\boldsymbol{\kappa} \in \{1,\dots,K_l\}^d$. Furthermore under periodic boundary conditions, this fixed kernel position implies that each position $\boldsymbol{\alpha}$ in $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}}$ originates from a different position $\boldsymbol{\alpha}'$ in the tensor $\mathbf{v}_{\boldsymbol{\alpha}',\mathrm{c}}$. Therefore the index mapping $f: \boldsymbol{\alpha} \to \boldsymbol{\alpha}'$ from $\{1,\dots,n\}^d$ to $\{1,\dots,n\}^d$ is bijective. We then have $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} = \mathbf{v}_{f(\boldsymbol{\alpha}),\mathrm{c}} \overset{\boldsymbol{\alpha}}{\sim} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ when $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}}$ and $\mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ are seen as random variables which depend on $\boldsymbol{\alpha}$ and $\mathbf{v}$ is given. In turn, this implies that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \overset{\mathbf{v},\boldsymbol{\alpha}}{\sim} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$, when they are seen as random variables which depend on $\mathbf{v}$, $\boldsymbol{\alpha}$.

□

Proposition 2. *If convolutions have periodic boundary conditions and the global spatial extent n is constant, then RF is statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathbf{s}^{l-1}$.*

Proof. This follows immediately from Lemma 1.

□

Corollary 3. *For any* c *and* $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$, *we have* $\mathbf{r}_\mathrm{f}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \overset{\mathbf{x},\boldsymbol{\alpha}}{\sim} \mathbf{f}_\mathrm{m}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$ *and* $\mathbf{r}_\mathrm{f}(\mathbf{s}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \overset{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}{\sim} \mathbf{f}_\mathrm{m}(\mathbf{s}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$. *Since the cardinality $|\mathcal{I}^l_\mathrm{c}| = K_l^d$ is the same for all channels* c, *it follows that*

$$\nu_2(\mathbf{x}^{l-1}) = \frac{1}{N_{l-1}} \operatorname{Tr} \boldsymbol{G}[\mathbf{f}_\mathrm{m}(\mathbf{x}^{l-1},\boldsymbol{\alpha})] = \frac{1}{R_l} \operatorname{Tr} \boldsymbol{G}[\mathbf{r}_\mathrm{f}(\mathbf{x}^{l-1},\boldsymbol{\alpha})],$$

$$\mu_2(\mathbf{x}^{l-1}) = \frac{1}{N_{l-1}} \operatorname{Tr} \boldsymbol{C}[\mathbf{f}_\mathrm{m}(\mathbf{x}^{l-1},\boldsymbol{\alpha})] = \frac{1}{R_l} \operatorname{Tr} \boldsymbol{C}[\mathbf{r}_\mathrm{f}(\mathbf{x}^{l-1},\boldsymbol{\alpha})].$$

C.2 Relaxing the assumptions on boundary conditions and constant spatial extent

In this section, we detail possible relaxations of the assumptions on boundary conditions and constant spatial extent n. The global spatial extent is denoted n_l when it is not constant.

C.2.1 Case of stationary inputs

Periodic extension. The *periodic extension* $\tilde{\mathbf{v}}$ of a random tensor $\mathbf{v} \in \mathbb{R}^{n\times\cdots\times n\times N}$ is defined as

$$\tilde{\mathbf{v}}_{\boldsymbol{\alpha}_1+k_1 n,\dots,\boldsymbol{\alpha}_d+k_d n,\mathrm{c}} = \mathbf{v}_{\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_d,\mathrm{c}},$$

with $(k_1,\dots,k_d) \in \mathbb{Z}^d$, and where $\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_d$ are the d components of the spatial position $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_d) \in \{1,\dots,n\}^d$.

Stationarity. The distribution of a random vector $\tilde{\mathbf{v}}$ is defined as *stationary* if, for any k and any configuration of spatial positions $(\boldsymbol{\alpha}_1,\dots,\boldsymbol{\alpha}_k)$ and channels $(\mathrm{c}_1,\dots,\mathrm{c}_k)$, the joint distribution of $\tilde{\mathbf{v}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_1,\mathrm{c}_1},\dots,\tilde{\mathbf{v}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_k,\mathrm{c}_k}$ is the same for all $\boldsymbol{\alpha}$.

Lemma 4. *If convolutions have periodic boundary conditions and the inputs* $\mathbf{x}$ *and* $\mathbf{s}$ *have stationary periodic extensions* $\tilde{\mathbf{x}}$ *and* $\tilde{\mathbf{s}}$*, then their periodic extension* $\tilde{\mathbf{x}}^l$ *and* $\tilde{\mathbf{s}}^l$ *remain stationary during propagation.*

Proof. First note that convolutions with periodic boundary conditions followed by periodic extensions on $\mathbf{v}$ are equivalent to convolutions on $\tilde{\mathbf{v}}$. This is also the case for componentwise operations such as batch normalization, nonlinear activation and their derivatives. So we can restrict our attention to periodic extensions.

Consider a given spatial shift $\boldsymbol{\alpha}$ and define the translation operator $T_{\boldsymbol{\alpha}}$, such that $T_{\boldsymbol{\alpha}}(\tilde{\mathbf{u}}) = \tilde{\mathbf{v}}$ with $\forall \boldsymbol{\alpha}', \mathrm{c}: \ \tilde{\mathbf{v}}_{\boldsymbol{\alpha}',\mathrm{c}} = \tilde{\mathbf{u}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}',\mathrm{c}}$. It is easy to see that $T_{\boldsymbol{\alpha}}$ commutes with convolutions as well as componentwise operations such as batch normalization, nonlinear activation and their derivatives. It follows that $T_{\boldsymbol{\alpha}}$ commutes with the input-output mapping Φ^l defined as $(\tilde{\mathbf{x}}^l, \tilde{\mathbf{s}}^l) = \Phi^l(\tilde{\mathbf{x}}, \tilde{\mathbf{s}})$, and thus we have

$$T_{\boldsymbol{\alpha}}(\tilde{\mathbf{x}}^l, \tilde{\mathbf{s}}^l) = \Phi^l(T_{\boldsymbol{\alpha}}(\tilde{\mathbf{x}}, \tilde{\mathbf{s}})), \tag{34}$$

where we adopted the notation $T_{\boldsymbol{\alpha}}(\tilde{\mathbf{u}}, \tilde{\mathbf{v}}) = \big(T_{\boldsymbol{\alpha}}(\tilde{\mathbf{u}}), T_{\boldsymbol{\alpha}}(\tilde{\mathbf{v}})\big)$. Now consider a given k and configuration of spatial positions $(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k)$ and channels $(\mathrm{c}_1, \ldots, \mathrm{c}_k)$ in $\tilde{\mathbf{x}}^l$ and $\tilde{\mathbf{s}}^l$. Due to limited convolutional spatial extent and due to Eq. (34), there exist a function Φ and a configuration of spatial positions $(\boldsymbol{\alpha}'_1, \ldots, \boldsymbol{\alpha}'_{k'})$ and channels $(\mathrm{c}'_1, \ldots, \mathrm{c}'_{k'})$ in $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{s}}$, such that we can write

$$\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}), \tag{35}$$

$$\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}), \tag{36}$$

$$\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{s}}^l_{\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \tilde{\mathbf{s}}_{\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}, \tilde{\mathbf{s}}_{\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}), \tag{37}$$

$$\tilde{\mathbf{s}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{s}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \tilde{\mathbf{s}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}, \tilde{\mathbf{s}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}). \tag{38}$$

By stationarity of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{s}}$, the terms on the right-hand sides of Eq. (35), Eq. (36) have the same distribution, and the terms on the right-hand sides of Eq. (37), Eq. (38) have the same distribution. It follows that the terms on the left-hand sides of both pairs of equations have the same distribution, meaning that $\tilde{\mathbf{x}}^l$ and $\tilde{\mathbf{s}}^l$ are stationary.

□

Lemma 5. *If convolutions have periodic boundary conditions, then* RF *is statistics-preserving with respect to any input* $\mathbf{v}$ *which has stationary periodic extension* $\tilde{\mathbf{v}}$ *for any global spatial extent* n_l*.*

Proof. If $\tilde{\mathbf{v}}$ has stationary distribution, it means in particular that for any channel c, the distribution of $\mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ is the same for all $\boldsymbol{\alpha} \in \{1, \ldots, n_{l-1}\}^d$. Fix a channel c and an index $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$ corresponding to a given convolution kernel position $\boldsymbol{\kappa} \in \{1, \ldots, K_l\}^d$. A given position $\boldsymbol{\alpha}$ in the receptive field tensor $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}}$ then corresponds to a given position $\boldsymbol{\alpha}'$ in the original tensor, such that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} = \mathbf{v}_{\boldsymbol{\alpha}',\mathrm{c}}$. Since the distribution of $\mathbf{v}_{\boldsymbol{\alpha}',\mathrm{c}}$ does not depend on $\boldsymbol{\alpha}'$, it follows that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \overset{\mathbf{v},\boldsymbol{\alpha}'}{\sim} \mathbf{v}_{\boldsymbol{\alpha}',\mathrm{c}}$ for given $\boldsymbol{\alpha}$ and random $\boldsymbol{\alpha}'$, and thus that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \overset{\mathbf{v},\boldsymbol{\alpha}}{\sim} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ for random $\boldsymbol{\alpha}$.

□

Proposition 6. *If convolutions have periodic boundary conditions and the input* $\mathbf{x}$ *has stationary periodic extension* $\tilde{\mathbf{x}}$*, then* RF *is statistics-preserving with respect to* $\mathbf{x}^l$ *and* $\mathbf{s}^l$*, for any global spatial extent* n_l*.*

Proof. This follows from Lemmas 4 and 5, and from the fact that the input sensitivity tensor has stationary periodic extension $\tilde{\mathbf{s}}$ due to its definition as a white noise tensor with independent and identically distributed components.

□

C.2.2 Case $n_l \gg K_l$

Proposition 7. *If the convolution stride is one in most layers (i.e.* $n_{l-1} = n_l$ *in most layers) and the global spatial extent is much larger than the convolutional spatial extent* $n_l \gg K_l$ *in most layers, then* RF *is approximately statistics-preserving with respect to* $\mathbf{x}^{l-1}$ *and* $\mathbf{s}^{l-1}$*, for any boundary conditions.*

Proof. Fix a layer $l-1$ such that $n_{l-1} = n_l$ and $n_l \gg K_l$. Denote $RF^{(\mathrm{p})}$ the receptive field mapping at layer l associated with periodic boundary conditions. Since $n_{l-1} = n_l \gg K_l$ the receptive fields $RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$, $RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:}$ and $RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$, $RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:}$ do not intersect boundary regions for most $\boldsymbol{\alpha}$, implying that $RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$, $RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:} = RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:}$ for most $\boldsymbol{\alpha}$. If we denote with F the cumulative distribution functions of any random variable, this implies for any index i_{c} that $F_{\mathbf{x},\boldsymbol{\alpha}}\big[RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] \simeq F_{\mathbf{x},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big]$ and $F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] \simeq F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big]$.

Since $RF^{(\mathrm{p})}$ is statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathbf{s}^{l-1}$ by Lemma 1, it follows that for any channel c and index $i_{\mathrm{c}} \in \mathcal{I}^l_{\mathrm{c}}$, we have $F_{\mathbf{x},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] = F_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\big]$ and $F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] = F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[\mathbf{s}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\big]$. We then deduce that $F_{\mathbf{x},\boldsymbol{\alpha}}\big[RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] \simeq F_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\big]$ and $F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\big] \simeq F_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[\mathbf{s}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\big]$, meaning that RF is approximately statistics-preserving for $\mathbf{x}^{l-1}$ and $\mathbf{s}^{l-1}$.

□

D Definition and Properties of the Normalized Sensitivity

D.1 Equivalence with previous definitions

In the fully-connected case $n = 1$, Philipp & Carbonell (2018) recently introduced the following coefficient:

$$\left(\frac{\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right]}{N_l}\frac{\mathbb{E}_{\mathrm{c}}\big[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^0_{\mathrm{c}}]\big]}{\mathbb{E}_{\mathrm{c}}\big[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\big]}\right)^{1/2}. \tag{39}$$

Let us prove the equivalence between the definitions of Eq. (3) and Eq. (39). In the fully-connected case, the spatial position $\boldsymbol{\alpha}$ can be ignored and tensors and feature map vectors coincide as $\mathbf{x}^l = \mathbf{f}_{\mathrm{m}}(\mathbf{x}^l)$, $\mathbf{s}^l = \mathbf{f}_{\mathrm{m}}(\mathbf{s}^l) = \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{s}^l)$. When $\mathbf{x}$ is fixed, the input-output Jacobian $\boldsymbol{J}^l = \frac{\partial \mathbf{x}^l}{\partial \mathbf{x}^0} \in \mathbb{R}^{N_l \times N_0}$ directly summarizes the propagation of the noise $\boldsymbol{\epsilon}^l = \boldsymbol{J}^l\boldsymbol{\epsilon}$, and thus the sensitivity $\mathbf{s}^l = \boldsymbol{J}^l\mathbf{s}$. Due to the white noise property $\mathbb{E}_{\mathbf{s}}\left[\mathbf{s}_i\mathbf{s}_j\right] = \delta_{ij}$,

$$\mathbb{E}_{\mathbf{x},\mathbf{s}}\Big[\sum_{\mathrm{c}}(\mathbf{s}^l_{\mathrm{c}})^2\Big] = \mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right], \quad \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\big[(\mathbf{s}^l_{\mathrm{c}})^2\big] = \frac{1}{N_l}\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right].$$

Here we clearly see the advantage of the sensitivity tensor to encode information on the Jacobian while avoiding increased dimensionality. Going back to our calculation, the definitions

$$\mu_2(\mathbf{s}^l) = \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{s}^l)^2_{\mathrm{c}}\big] = \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\big[\mathbf{f}_{\mathrm{m}}(\mathbf{s}^l)^2_{\mathrm{c}}\big] = \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\big[(\mathbf{s}^l_{\mathrm{c}})^2\big],$$
$$\mu_2(\mathbf{x}^l) = \mathbb{E}_{\mathbf{x},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l)^2_{\mathrm{c}}\big] = \mathbb{E}_{\mathrm{c}}\big[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\big],$$

finally give the equivalence between the two definitions:

$$\zeta^l = \left(\frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)}\right)^{1/2} = \left(\frac{\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right]}{N_l}\frac{\mathbb{E}_{\mathrm{c}}\big[\mathrm{Var}_{\mathbf{x}}\big[\mathbf{x}^0_{\mathrm{c}}\big]\big]}{\mathbb{E}_{\mathrm{c}}\big[\mathrm{Var}_{\mathbf{x}}\big[\mathbf{x}^l_{\mathrm{c}}\big]\big]}\right)^{1/2}.$$

Philipp & Carbonell (2018) chose the terminology of *nonlinearity coefficient* for this metric. While our analysis unveils a strong relationship between ζ^l and the nonlinearity ϕ, it also reveals a strong relationship with batch normalization which is still a linear operation. So we chose instead the terminology of normalized sensitivity.

D.2 Property of Normalized Sensitivity

Proposition 8. *The sensitivity tensor and the tensor $\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ containing for given $\boldsymbol{\alpha}, \mathrm{c}$ the derivatives of $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ with respect to $\mathbf{x}$ are related by $\mathbb{E}_{\mathbf{s}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = ||\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2$.*

Proof of Proposition 8. Due to the definition of $\boldsymbol{\epsilon}^l$ as a small corruption to $\mathbf{x}^l$, $\boldsymbol{\epsilon}^l_{\boldsymbol{\alpha},\mathrm{c}}$ can be written as a function of $\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and the input noise $\boldsymbol{\epsilon}$,

$$\boldsymbol{\epsilon}^l_{\boldsymbol{\alpha},\mathrm{c}} = \langle \nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}, \boldsymbol{\epsilon} \rangle,$$

where $\langle\,.\,\rangle$ denotes the standard dot product in input space. It follows from the definition of $\mathbf{s}^l = \boldsymbol{\epsilon}^l \,/\, \sigma_{\boldsymbol{\epsilon}}$ that $\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}} = \langle \nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}, \mathbf{s} \rangle$. Due to the white noise property $\mathbb{E}_{\mathbf{s}}\left[\mathbf{s}_i\mathbf{s}_j\right] = \delta_{ij}$, we then get

$$\mathbb{E}_{\mathbf{s}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = ||\nabla_{\mathbf{x}}\mathbf{x}_{\boldsymbol{\alpha},\mathrm{c}}||_2^2.$$

□

Proposition 9. *Define* $\mathbf{v}$ *the rescaling by a constant factor of* $\mathbf{x}$ *with unit variance* $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v},\boldsymbol{\alpha})^2_{\mathrm{c}}] = 1$*, and* $\mathbf{v}^l$ *the rescaling by a constant factor of* $\mathbf{x}^l$ *with unit variance* $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}] = 1$*. When* $\mathbf{v}^l$ *is considered as a function of* $\mathbf{v}$*,* ζ^l *measures an expected sensitivity of* $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})$ *as* $\zeta^l = \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{v}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right]^{1/2}$.

Proof of Proposition 9. First we work with the non-normalized $\mathbf{x}$ and $\mathbf{x}^l$. We can express the second-order central moment of $\mathbf{s}$ as

$$\begin{aligned}
\mu_2(\mathbf{s}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\mathbb{E}_{\mathbf{s}}\left[\mathbf{f}_{\mathrm{m}}(\mathbf{s}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\mathbb{E}_{\mathbf{s}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] && (40)\\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right] && (41)\\
&= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\mathbf{f}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right],
\end{aligned}$$

where Eq. (40) follows from $\mathbf{s}^l$ being centered and Eq. (41) follows from Proposition 8. Now for given c and $\boldsymbol{\alpha}$, the difference $\mathbf{f}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}} - \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}$ is constant with respect to $\mathbf{x}^l$, and thus with respect to $\mathbf{x}$. The derivatives of $\mathbf{f}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}$ and $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}$ with respect to $\mathbf{x}$ are then equal for all $\mathbf{x}$, $\boldsymbol{\alpha}$ and c. Together with the definition of $\mu_2(\mathbf{x}^l)$, we deduce the following:

$$\begin{aligned}
\mu_2(\mathbf{s}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right],\\
\mu_2(\mathbf{x}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right],\\
\zeta^l &= \left(\frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)}\right)^{1/2} = \left(\frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right]\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x},\boldsymbol{\alpha})^2_{\mathrm{c}}\right]}{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right]}\right)^{1/2}. \qquad (42)
\end{aligned}$$

Defining $\mathbf{v}^l = \mathbf{x}^l \,/\, \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right]^{1/2}$, $\mathbf{v} = \mathbf{x} \,/\, \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x},\boldsymbol{\alpha})^2_{\mathrm{c}}\right]^{1/2}$, we get $\nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha}) = \nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha}) \,/\, \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right]^{1/2}$ and $\nabla_{\mathbf{v}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha}) = \nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})\,\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x},\boldsymbol{\alpha})^2_{\mathrm{c}}\right]^{1/2}$. It follows that

$$\begin{aligned}
\mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right] &= 1,\\
\mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v},\boldsymbol{\alpha})^2_{\mathrm{c}}\right] &= 1,\\
\mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{v}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right] &= \frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right]\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x},\boldsymbol{\alpha})^2_{\mathrm{c}}\right]}{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{x}^l,\boldsymbol{\alpha})^2_{\mathrm{c}}\right]}. \qquad (43)
\end{aligned}$$

Finally combining Eq. (42) and Eq. (43), we get $\zeta^l = \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{v}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{v}^l,\boldsymbol{\alpha})_{\mathrm{c}}||_2^2\right]^{1/2}$.

□

Illustration. Let us consider the case of fully-connected networks with L layers, and with 1-dimensional input $N_0 = 1$ and 1-dimensional output $N_L = 1$. Denote $\mathbf{x}^L = \Phi(\mathbf{x})$ the original input-output mapping and $\mathbf{v}^L = \tilde{\Phi}(\mathbf{v})$ the rescaled input-output mapping. Proposition 9 then simply becomes $\zeta^L = \mathbb{E}_{\mathbf{v}}\left[\tilde{\Phi}'(\mathbf{v})^2\right]^{1/2}$. In Fig. 4 we show the resut of the propagation $\mathbf{v}^L = \Phi(\mathbf{v})$ of an input $\mathbf{v}$ having a mixture of Gaussians distribution in three different cases:

– In Fig. 4a the input $\mathbf{v}$ is propagated through $L = 1$ layer with sigmoid activation. After propagation, the expected derivative is low since each mode of the Gaussian mixture appears in a relatively flat region of the sigmoid activation. This is clearly shown by the input-output data points and by the histogram of the outputs.

– In Fig. 4b the input $\mathbf{v}$ is propagated through $L = 1$ layer with linear activation.

– In Fig. 4c the input $\mathbf{v}$ is propagated through $L = 25$ randomly initialized layers, with ReLU activation in the $L - 1$ first layers and linear activation in the final layer. After propagation, the expected derivative is high due to the erratic behavior of the input-output mapping.

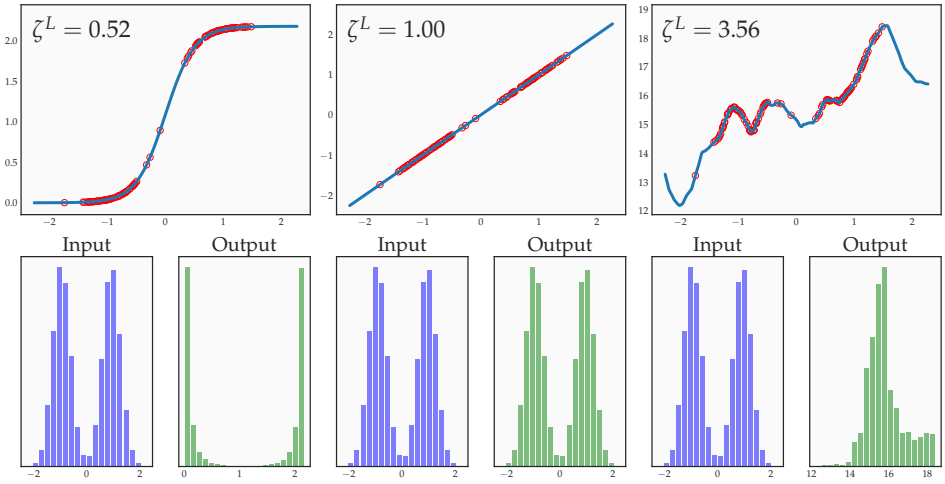

Figure 4: Illustration of the normalized sensitivity for fully-connected networks of L layers, with 1-dimensional input $N_0 = 1$ and 1-dimensional output $N_L = 1$. The distribution of the input $\mathbf{v}$ is a mixture of two Gaussians. We show the result of the propagation in three different cases: *(a)* $L = 1$ layer with sigmoid activation, *(b)* $L = 1$ layer with linear activation, *(c)* $L = 25$ randomly initialized layers, with $N_l = 100$ channels and ReLU activation for $1 \leq l < L$, and linear activation in the final layer $l = L$. *Top*: full input-output mapping (blue curve) and randomly sampled input-output data points (red circles). *Bottom*: histograms of inputs and outputs.

E Moments of Vanilla Networks

E.1 Lemma on the sum of increments

Lemma 10. *Let X_k be a sequence of random variables which depend on Θ^k, and denote $Y_k = \mathbb{E}_{\theta^k}[X_k]$ and $Z_k = X_k - \mathbb{E}_{\theta^k}[X_k]$. If there exist constants $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$ with $\forall k \leq l$: $m_{\min} \leq Y_k \leq m_{\max}$ and $v_{\min} \leq \mathrm{Var}_{\Theta^k}[Z_k] \leq v_{\max}$, then:*

(i) The increments Z_k are centered and non-correlated:

$$\forall k \leq l : \ \mathbb{E}_{\Theta^k}[Z_k] = 0, \qquad \forall k \neq k' \leq l : \ \mathbb{E}_{\Theta^{\max(k,k')}}[Z_k Z_{k'}] = 0.$$

(ii) There exist random variables m_l and s_l such that s_l is centered and

$$\sum_{k=1}^{l} X_k = l m_l + \sqrt{l} s_l, \quad m_{\min} \leq m_l \leq m_{\max}, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l}[s_l] \leq v_{\max}.$$

Proof of (i). First we show that Z_k is centered:

$$\begin{aligned} \mathbb{E}_{\theta^k}[Z_k] &= \mathbb{E}_{\theta^k}[X_k] - \mathbb{E}_{\theta^k}[X_k] = 0, \\ \mathbb{E}_{\Theta^k}[Z_k] &= \mathbb{E}_{\Theta^{k-1}}[\mathbb{E}_{\theta^k}[Z_k]] = 0. \end{aligned} \tag{44}$$

Now consider $k < k'$, which implies $k \leq k' - 1$ and thus that Z_k is a random variable which only depends on $\Theta^{k'-1}$. Then we can write

$$\begin{aligned}\mathbb{E}_{\Theta^{k'}}[Z_k Z_{k'}] &= \mathbb{E}_{\Theta^{k'-1}}\mathbb{E}_{\theta^{k'}}[Z_k Z_{k'}] \\ &= \mathbb{E}_{\Theta^{k'-1}}\left[Z_k\, \mathbb{E}_{\theta^{k'}}[Z_{k'}]\right] \\ &= 0, \end{aligned} \tag{45}$$

where Eq. (45) follows from Eq. (44).

□

Proof of (ii). Denote $M_l = \sum_{k=1}^{l} Y_k$ and $S_l = \sum_{k=1}^{l} Z_k$. We then have

$$\begin{aligned}\mathbb{E}_{\Theta^l}[S_l] &= \sum\nolimits_k \mathbb{E}_{\Theta^l}[Z_k] = 0, \\ \mathbb{E}_{\Theta^l}[S_l^2] &= \sum\nolimits_{k,k'} \mathbb{E}_{\Theta^l}[Z_k Z_{k'}], \\ \mathrm{Var}_{\Theta^l}[S_l] &= \sum\nolimits_k \mathbb{E}_{\Theta^k}[Z_k^2] = \sum\nolimits_k \mathrm{Var}_{\Theta^k}[Z_k], \end{aligned} \tag{46}$$

where Eq. (46) follows from (i). The hypothesis then gives $l m_{\min} \leq M_l \leq l m_{\max}$ and $l v_{\min} \leq \mathrm{Var}_{\Theta^l}[S_l] \leq l v_{\max}$. If we define $m_l = M_l \,/\, l$ and $s_l = S_l \,/\, \sqrt{l}$, then s_l is centered and the telescoping sum $\sum_{k=1}^{l} X_k = \sum_{k=1}^{l} Y_k + \sum_{k=1}^{l} Z_k$ can be written as required:

$$\sum_{k=1}^{l} X_k = M_l + S_l = l m_l + \sqrt{l} s_l, \quad m_{\min} \leq m_l \leq m_{\max}, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l}[s_l] \leq v_{\max}.$$

□

E.2 Proof of Theorem 1

Theorem 1. Moments of vanilla networks. *Denote A_l the event $\{\nu_2(\mathbf{x}^l) > 0\}$ and A'_l the complementary event $\{\nu_2(\mathbf{x}^l) = 0\} = \{\mathbb{P}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0] = 1\}$. Then:*

(i) $\prod_{k=1}^{l}\left(1 - 2^{-N_k}\right) \leq \mathbb{P}[A_l] \leq \prod_{k=1}^{l}\left(1 - 2^{-K_k^d N_{k-1} N_k}\right)$

(ii) There exist positive constants $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max} > 0$ and sequences of random variables $(m_l), (m'_l), (s_l), (s'_l)$ such that under A_l, s_l, s'_l are centered and

$$\begin{aligned}&\log \nu_2(\mathbf{x}^l) = -l m_l + \sqrt{l} s_l + \log \nu_2(\mathbf{x}^0), && m_{\min} \leq m_l \leq m_{\max}, && v_{\min} \leq \mathrm{Var}_{\Theta^l | A_l}[s_l] \leq v_{\max}, \\ &\log \mu_2(\mathbf{s}^l) = -l m'_l + \sqrt{l} s'_l, && m_{\min} \leq m'_l \leq m_{\max}, && v_{\min} \leq \mathrm{Var}_{\Theta^l | A_l}[s'_l] \leq v_{\max}. \end{aligned}$$

Proof of (i). We use the definitions and notations from section B. We further denote $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \ldots, \lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{G}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$, and $\hat{\boldsymbol{W}}^l = \boldsymbol{W}^l(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$. We then get

$$\begin{aligned}\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^l_{\mathrm{c},:}\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right)^2\right] \\ &= \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \lambda_i. \end{aligned} \tag{47}$$

Given the assumption of standard initialization, biases are initialized with zeros and weights are initialized as $\boldsymbol{W}^l \overset{\theta^l}{\sim} \hat{\boldsymbol{W}}^l \overset{\theta^l}{\sim} \mathcal{N}(0, \sqrt{2 \,/\, R_l}\boldsymbol{I})$. Under A_{l-1}, it follows from Corollary 3 that $\mathrm{Tr}\, \boldsymbol{G}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})] = R_l \nu_2(\mathbf{x}^{l-1}) > 0$ and thus $\mathrm{Tr}\, \boldsymbol{G}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})] = \sum_i \lambda_i > 0$. Combined with Eq. (47), we get

$$\forall \mathrm{c}: \ \mathbb{P}_{\theta^l | A_{l-1}}\left[\nu_{2,\mathrm{c}}(\mathbf{y}^l) = 0\right] = 0. \tag{48}$$

Now we introduce the symmetric propagation from section B and we denote $A_{l,\mathrm{c}} = \{\nu_{2,\mathrm{c}}(\mathbf{x}^l) \neq 0\}$, $\tilde{A}_{l,\mathrm{c}} = \{\nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) \neq 0\}$ and $A'_{l,\mathrm{c}} = \{\nu_{2,\mathrm{c}}(\mathbf{x}^l) = 0\}$, $\tilde{A}'_{l,\mathrm{c}} = \{\nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = 0\}$ the complementary events. It follows from Eq. (48) and Eq. (25) that $\mathbb{P}_{\theta^l | A_{l-1}}[A'_{l,\mathrm{c}} \cap \tilde{A}'_{l,\mathrm{c}}] = 0$. Furthermore by spherical symmetry of the propagation, $\nu_{2,\mathrm{c}}(\mathbf{x}^l) \overset{\theta^l}{\sim} \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l)$ and thus $\mathbb{P}_{\theta^l | A_{l-1}}[A'_{l,\mathrm{c}}] = \mathbb{P}_{\theta^l | A_{l-1}}[\tilde{A}'_{l,\mathrm{c}}]$. We

then get

$$\forall \mathrm{c}: \ \mathbb{P}_{\theta^l|A_{l-1}}\left[A'_{l,\mathrm{c}} \cup \tilde{A}'_{l+1,\mathrm{c}}\right] = \mathbb{P}_{\theta^l|A_{l-1}}\left[A'_{l,\mathrm{c}}\right] + \mathbb{P}_{\theta^l|A_{l-1}}\left[\tilde{A}'_{l+1,\mathrm{c}}\right] \leq 1,$$
$$\forall \mathrm{c}: \ \mathbb{P}_{\theta^l|A_{l-1}}\left[A'_{l,\mathrm{c}}\right] \leq \frac{1}{2}.$$

Since $A'_l = \bigcap_{\mathrm{c}} A'_{l,\mathrm{c}}$ and since the events $A'_{l,\mathrm{c}}$ are independent conditionally on Θ^{l-1} and A_{l-1}, we conclude that $\mathbb{P}_{\theta^l|A_{l-1}}[A'_l] = \prod_{\mathrm{c}} \mathbb{P}_{\theta^l|A_{l-1}}[A'_{l,\mathrm{c}}] \leq 2^{-N_l}$.

The other side of the inequality is easier. If $\forall \mathrm{c}, i_{\mathrm{c}}$: $\boldsymbol{W}^l_{\mathrm{c},i_{\mathrm{c}}} \leq 0$, it follows that $\forall \mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$: $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} < 0$, $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} = (\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^+ = 0$, and thus $\nu_2(\mathbf{x}^l) = 0$. Therefore $\mathbb{P}_{\theta^l|A_{l-1}}[A'_l] \geq 2^{-K_l^d N_{l-1} N_l}$, since $K_l^d N_{l-1} N_l = R_l N_l$ is the number of elements in $\boldsymbol{W}^l$. We finally get

$$1 - 2^{-N_l} \leq \mathbb{P}_{\theta^l|A_{l-1}}\left[A_l\right] \leq 1 - 2^{-K_l^d N_{l-1} N_l}. \tag{49}$$

Since $\mathbb{P}[A_0] = 1$, due to the assumption $\nu_2(\mathbf{x}) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^2_{\boldsymbol{\alpha},\mathrm{c}}] > 0$, it follows that $1 - 2^{-N_1} \leq \mathbb{P}[A_1] \leq 1 - 2^{-K_1^d N_0 N_1}$. This proves (i) for $l = 1$.

Now we proceed by induction and suppose that (i) is true for given $l - 1$. Using Eq. (49), we get

$$\prod_{k=1}^{l} \left(1 - 2^{-N_k}\right) \leq \mathbb{P}_{\Theta^l}\left[A_l\right] = \mathbb{P}_{\Theta^{l-1}}\left[A_{l-1}\right] \mathbb{P}_{\theta^l|A_{l-1}}\left[A_l\right] \leq \prod_{k=1}^{l} \left(1 - 2^{-K_k^d N_{k-1} N_k}\right),$$

meaning that (i) is true for l. (i) is thus true for all l.

□

Proof of (ii) for $\nu_2(\mathbf{x}^l)$. Again we denote $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \dots, \lambda_{R_l})$ the eigenvectors and eigenvalues of $\boldsymbol{G}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$, and $\hat{\boldsymbol{W}}^l = \boldsymbol{W}^l(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$. We further define

$$u^l_{\mathrm{c}} = \begin{cases} \frac{\nu_{2,\mathrm{c}}(\mathbf{x}^l)}{\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l)} & \text{if } \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) > 0 \\ 0 & \text{otherwise.} \end{cases}$$

Combining the definition of u^l_{c} with Eq. (25) and Eq. (47), we get

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) = u^l_{\mathrm{c}}\, \nu_{2,\mathrm{c}}(\mathbf{y}^l),$$
$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) = u^l_{\mathrm{c}} \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \lambda_i = R_l\, \nu_2(\mathbf{x}^{l-1})\, u^l_{\mathrm{c}} \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i, \tag{50}$$

where we defined $\hat{\lambda}_i = \lambda_i \,/\, \sum_j \lambda_j$ and used $\sum_j \lambda_j = \operatorname{Tr} \boldsymbol{G}\left[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right] = R_l \nu_2(\mathbf{x}^{l-1})$ due to Corollary 3. The symmetric propagation gives

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = R_l\, \nu_2(\mathbf{x}^{l-1})\, (1 - u^l_{\mathrm{c}}) \sum\nolimits_i \left(-\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i,$$
$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = R_l\, \nu_2(\mathbf{x}^{l-1}) \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i. \tag{51}$$

By symmetry of the propagation $\nu_{2,\mathrm{c}}(\mathbf{x}^l) \overset{\theta^l}{\sim} \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l)$. Combined with Eq. (51) and the assumption of standard initialization, we deduce

$$\begin{aligned} 2\mathbb{E}_{\theta^l|A_{l-1}}\left[\nu_{2,\mathrm{c}}(\mathbf{x}^l)\right] &= \mathbb{E}_{\theta^l|A_{l-1}}\left[\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l)\right] \\ &= \mathbb{E}_{\theta^l|A_{l-1}}\left[R_l\, \nu_2(\mathbf{x}^{l-1}) \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i\right] \\ &= R_l\, \nu_2(\mathbf{x}^{l-1}) \frac{2}{R_l} \sum\nolimits_i \hat{\lambda}_i = 2\nu_2(\mathbf{x}^{l-1}). \end{aligned}$$

We then obtain $\forall \mathrm{c}: \ \mathbb{E}_{\theta^l|A_{l-1}}[\nu_{2,\mathrm{c}}(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1})$, and thus

$$\mathbb{E}_{\theta^l|A_{l-1}}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1}), \qquad \mathbb{E}_{\theta^l|A_{l-1}}[\delta\nu_2(\mathbf{x}^l)] = 1. \tag{52}$$

Both $\log x$ and $(\log x)^2$ are integrable at zero since $\int \log x\, dx = x\log x - x$ and $\int (\log x)^2 dx = x(\log x)^2 - 2x\log x + 2x$. Combined with Eq. (50), we deduce that $\log \delta\nu_2(\mathbf{x}^l)$ has well-defined conditional expectation and variance under A_l. Furthermore by Eq. (49), A_l has probability exponentially low in N_l, which gives

$$|\mathbb{E}_{\theta^l|A_l}[\delta\nu_2(\mathbf{x}^{l+1})] - 1| \ll 1, \qquad \mathbb{E}_{\theta^l|A_l}[\log \delta\nu_2(\mathbf{x}^{l+1})] < 0, \tag{53}$$

where Eq. (53) is obtained by log-concavity. We now write the evolution of $\nu_2(\mathbf{x}^l)$ as

$$\log \nu_2(\mathbf{x}^l) - \nu_2(\mathbf{x}^0) = \sum_{k=1}^{l} \log \delta\nu_2(\mathbf{x}^k).$$

Let us define $X_k = \log \delta\nu_2(\mathbf{x}^k)$, $Y_k = \mathbb{E}_{\theta^k|A_l}[\log \delta\nu_2(\mathbf{x}^k)]$ and $Z_k = \log \delta\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k|A_l}[\log \delta\nu_2(\mathbf{x}^k)]$ and apply Lemma 10 for each l conditionally on A_l. Suppose there exist $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$, such that for each l we have conditionally on A_l that $\forall k \leq l$, $m_{\min} \leq -Y_k \leq m_{\max}$ and $v_{\min} \leq \mathrm{Var}_{\theta^k|A_k}[Z_k] \leq v_{\max}$ which implies $v_{\min} \leq \mathrm{Var}_{\Theta^k|A_l}[Z_k] \leq v_{\max}$. Then we have $\forall k \leq l$, $-m_{\max} \leq Y_k \leq -m_{\min}$ and by Lemma 10 there exist sequences of random variables (m_l) and (s_l) such that $\forall l$ under A_l, s_l is centered and

$$\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0) = l m_l + \sqrt{l} s_l, \quad -m_{\max} \leq m_l \leq -m_{\min}, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\max}.$$

By simply changing the variable m_l to $-m_l$, we get

$$\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0) = -l m_l + \sqrt{l} s_l, \quad m_{\min} \leq m_l \leq m_{\max}, \qquad v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\max}.$$

To obtain the bounds $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$, we consider extreme cases for u_{c}^l and $\sum_i (\hat{\boldsymbol{W}}_{\mathrm{c},i}^l)^2 \hat{\lambda}_i$ in Eq. (50). Denoting $\chi^2(N)$ the chi-square distribution with N degree of freedom, we obtain *minimum bounds* by considering $u_{\mathrm{c}}^l \overset{\theta^l}{\sim} 1/2$ and $\sum_i (\hat{\boldsymbol{W}}_{\mathrm{c},i}^l)^2 \hat{\lambda}_i \overset{\theta^l}{\sim} \chi^2(R_l)$. This leads to $\log \delta\nu_2(\mathbf{x}^l) \overset{\theta^l}{\sim} \chi^2(N_l R_l)$. Denoting $\mathrm{Bern}(1/2)$ the Bernouilli distribution with $p = 1/2$, we obtain *maximum bounds* by considering $u_{\mathrm{c}} \overset{\theta^l}{\sim} \mathrm{Bern}(1/2)$ and $\sum_i (\hat{\boldsymbol{W}}_{\mathrm{c},i}^l)^2 \hat{\lambda}_i \overset{\theta^l}{\sim} \chi^2(1)$. The conditionality on A_l has highly negligible impact in practice.

Let us give an example in the fully-connected case with constant width $N_l = 100$. We then find numerically $m_{\min} \simeq 9.7{\times}10^{-5}$ and $v_{\min} \simeq 2.0{\times}10^{-4}$ as minimum bounds and $m_{\max} \simeq 2.5{\times}10^{-2}$ and $v_{\max} \simeq 5.2{\times}10^{-2}$ as maximum bounds. The length scale is experimentally close to $L_{\max} = 1/m_{\max} \simeq 40$ for $\nu_2(\mathbf{x}^l)$.

□

Proof of (ii) for $\mu_2(\mathbf{s}^l)$. Let us denote $B_l = \left\{\forall k \leq l, \forall \mathrm{c} : \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\}$ and $B_l' = \left\{\exists k \leq l, \exists \mathrm{c} : \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) \neq \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\} = \bigcup_{k=1}^{l} \left\{\exists \mathrm{c} : \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) \neq \mu_{2,\mathrm{c}}(\mathbf{t}^k)\right\}$ the complementary event. Eq. (28) gives $\forall k$: $\mathbb{P}_{\theta^k}\left[\left\{\exists \mathrm{c} : \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) \neq \mu_{2,\mathrm{c}}(\mathbf{t}^k)\right\}\right] = 0$. Since B_l' is the union of probability zero events, it follows that $\mathbb{P}_{\theta^l}(B_l') = 0$, $\mathbb{P}_{\theta^l}(B_l) = 1$, and thus $\mathbb{P}_{\Theta^l}(B_l') = 0$, $\mathbb{P}_{\Theta^l}(B_l) = 1$.

Since B_l has probability 1, the conditionality on B_l leaves moments of random variables unchanged. To see this, consider the moment of order p of a random variable x:

$$\begin{aligned}
\left(\mathbb{E}_{\Theta^l|A_l}[x^p] - \mathbb{E}_{\Theta^l|A_l,B_l}[x^p]\right)^2 &= \left(\mathbb{E}_{\Theta^l|A_l}[x^p] - \frac{1}{\mathbb{P}_{\Theta^l|A_l}(B_l)} \mathbb{E}_{\Theta^l|A_l}[\mathbf{1}_{B_l} x^p]\right)^2 \\
&= \mathbb{E}_{\Theta^l|A_l}[\mathbf{1}_{B_l'} x^p]^2 \\
&\leq \mathbb{E}_{\Theta^l|A_l}[x^{2p}]\, \mathbb{P}_{\Theta^l|A_l}[B_l'] = 0,
\end{aligned} \tag{54}$$

where Eq. (54) is obtained with Cauchy-Schwarz inequality and the implicit assumption that x has well-defined moment of order $2p$. It follows that all arguments used in the proof of (i) remain valid, in particular regarding the distribution of $\boldsymbol{W}^l$. Conditionality on B_l, the proof then proceeds identically by simply replacing $\nu_2(\mathbf{x}^l)$ by $\mu_2(\mathbf{s}^l)$, $\mathbf{y}^l$ by $\mathbf{t}^l$, $\boldsymbol{G}$ by $\boldsymbol{C}$, and using the identity with μ_2 instead of ν_2 in Corollary 3. In particular, we have

$$\mathbb{E}_{\theta^l|A_{l-1},B_l}[\mu_2(\mathbf{s}^l)] = \mu_2(\mathbf{s}^{l-1}), \qquad \mathbb{E}_{\theta^l|A_{l-1},B_l}[\delta\mu_2(\mathbf{s}^l)] = 1. \tag{55}$$

Furthermore under $A_l \cap B_l$, for the same positive constants $m_{\min}, m_{\max}, v_{\min}, v_{\max} > 0$ as previously defined, there exist sequences of random variables (m'_l) and (s'_l) such that $\forall l$ under $A_l \cap B_l$, s'_l is centered and

$$\log \mu_2(\mathbf{s}^l) = -lm'_l + \sqrt{l}s'_l, \quad m_{\min} \leq m'_l \leq m_{\max}, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l,B_l}[s'_l] \leq v_{\max},$$

where we used the fact that $\mu_2(\mathbf{s}^0) = 1$ by the definition of $\mathbf{s}$. Now we extend m'_l and s'_l to B_l by setting m'_l to any value between $m_{\min}$ and $m_{\max}$ and s'_l such that $\log \mu_2(\mathbf{s}^l) = -lm'_l + \sqrt{l}s'_l$. The reasoning of Eq. (54) then ensures that $\mathbb{E}_{\Theta^l|A_l,B_l}[s'_l] = \mathbb{E}_{\Theta^l|A_l}[s'_l]$ and $\mathbb{E}_{\Theta^l|A_l,B_l}[(s'_l)^2] = \mathbb{E}_{\Theta^l|A_l}[(s'_l)^2]$, and $\mathrm{Var}_{\Theta^l|A_l,B_l}[s'_l] = \mathrm{Var}_{\Theta^l|A_l}[s'_l]$. It means that $\forall l$ under A_l, s'_l is centered and

$$\log \mu_2(\mathbf{s}^l) = -lm'_l + \sqrt{l}s'_l, \quad m_{\min} \leq m'_l \leq m_{\max}, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l}[s'_l] \leq v_{\max}.$$

□

E.3 Relation to the terms $\overline{m}$, $\underline{m}$, $\underline{o}$ defined in section 4

Here we relate Theorem 1 to the terms $\overline{m}$, $\underline{m}$, $\underline{s}$ defined in section 4. By Eq. (52), $|\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \ll 1$, and thus:

$$|\overline{m}[\nu_2(\mathbf{x}^k)]| = |\log \mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)]| \simeq |\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \ll 1.$$

As in the proof of Theorem 1, we denote $B_l = \left\{\forall k \leq l,\ \forall \mathrm{c}:\ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) = \mu_{2,\mathrm{c}}(\mathbf{t}^l)\right\}$ and B'_l the complementary event. Then conditionally on A_k and B_k:

$$|\log \mathbb{E}_{\theta^k|A_k,B_k}[\delta\nu_2(\mathbf{x}^k)]| \simeq |\mathbb{E}_{\theta^k|A_k,B_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \ll 1.$$

The reasoning of Eq. (54) can be applied to $\delta\nu_2(\mathbf{x}^k)$, which results in $\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] = \mathbb{E}_{\theta^k|A_k,B_k}[\delta\nu_2(\mathbf{x}^k)]$. Therefore we also get

$$|\overline{m}[\mu_2(\mathbf{s}^k)]| = |\log \mathbb{E}_{\theta^k|A_k}[\delta\mu_2(\mathbf{s}^k)]| \ll 1.$$

The terms $\overline{m}[\nu_2(\mathbf{x}^k)]$ and $\overline{m}[\mu_2(\mathbf{s}^k)]$ are thus vanishing and the evolution is dominated by the terms $\overline{m}[\nu_2(\mathbf{x}^k)] < 0$, $\overline{m}[\mu_2(\mathbf{s}^k)] < 0$. These terms correspond to Y_k in the proof of Theorem 1 (ii).

E.4 Convergence in probability to zero

Corollary 11. *Conditionally on A_l the variables $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ still converge in probability to zero:*

$$\forall \epsilon:\ \mathbb{P}_{\Theta^l|A_l}\left[|\nu_2(\mathbf{x}^l)| > \epsilon\right] \to 0, \qquad \forall \epsilon:\ \mathbb{P}_{\Theta^l|A_l}\left[|\mu_2(\mathbf{s}^l)| > \epsilon\right] \to 0.$$

Proof. Consider a given ϵ and the evolution of $\nu_2(\mathbf{x}^l)$. From theorem 1 (ii), we can write under A_l: $\log \nu_2(\mathbf{x}^l) = -lm_l + \sqrt{l}s_l + \log \nu_2(\mathbf{x}^0)$, with s_l centered and $m_{\min} \leq m_l \leq m_{\max}$, $v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\max}$. Therefore:

$$\begin{aligned}
\mathbb{P}_{\Theta^l|A_l}\left[|\nu_2(\mathbf{x}^l)| > \epsilon\right] &= \mathbb{P}_{\Theta^l|A_l}\left[\log \nu_2(\mathbf{x}^l) > \log \epsilon\right] = \mathbb{P}_{\Theta^l|A_l}\left[\sqrt{l}s_l > lm_l + \log \epsilon - \log \nu_2(\mathbf{x}^0)\right] \\
&\leq \mathbb{P}_{\Theta^l|A_l}\left[s_l > \frac{1}{\sqrt{l}}\left(lm_l + \log \epsilon - \log \nu_2(\mathbf{x}^0)\right)\right] \\
&\leq \mathbb{P}_{\Theta^l|A_l}\left[|s_l| > \frac{1}{\sqrt{l}}\left(lm_{\min} + \log \epsilon - \log \nu_2(\mathbf{x}^0)\right)\right].
\end{aligned} \tag{56}$$

Chebyshev's inequality on the centered random variable s_l then gives:

$$\mathbb{P}_{\Theta^l|A_l}\left[|\nu_2(\mathbf{x}^l)| > \epsilon\right] \leq \frac{l}{\left(l m_{\min} + \log\epsilon - \log\nu_2(\mathbf{x}^0)\right)^2} \operatorname{Var}_{\Theta^l|A_l}[s_l]$$
$$\leq \frac{l}{\left(l m_{\min} + \log\epsilon - \log\nu_2(\mathbf{x}^0)\right)^2} v_{\max} \sim \frac{1}{l}\frac{v_{\max}}{m_{\min}^2}$$

where $\sim$ denotes the equivalence for large l. It follows that $\mathbb{P}_{\Theta^l|A_l}\left[|\nu_2(\mathbf{x}^l)| > \epsilon\right] \to 0$. The same analysis applied to $\mu_2(\mathbf{s}^l)$ gives $\mathbb{P}_{\Theta^l|A_l}\left[|\mu_2(\mathbf{s}^l)| > \epsilon\right] \to 0$.

□

F Normalized Sensitivity increments of Vanilla Networks

F.1 Proof of Theorem 2

Theorem 2. Normalized Sensitivity increments of vanilla networks. *Under A_{l-1}, the dominating term in the evolution of the normalized sensitivity is:*

$$\delta\zeta^l \simeq \exp\left(\overline{m}_{\text{vanilla}}\left[\zeta^l\right]\right) = \left(1 - \mathbb{E}_{\mathrm{c},\theta^l|A_{l-1}}\left[\frac{\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-1/2}, \tag{57}$$

where $\mathbf{y}^{l,+} = \max(\mathbf{y}^l, 0)$ *and* $\mathbf{y}^{l,-} = \max(-\mathbf{y}^l, 0)$.

Proof. The dominating term in the evolution of ζ^l is $\frac{1}{2}\left(\overline{m}[\mu_2(\mathbf{s}^l)] - \overline{m}[\mu_2(\mathbf{x}^l)\right)$. The terms $\overline{m}[\mu_2(\mathbf{s}^l)]$ and $\overline{m}[\mu_2(\mathbf{x}^l)$ are simply obtained by considering $\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{s}^l)]$ and $\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{x}^l)]$.

By Eq. (53) in the proof of Theorem 1: $\mathbb{E}_{\theta^l|A_{l-1},B_l}[\delta\mu_2(\mathbf{s}^l)] = 1$. By replicating the reasoning of Eq. (54), this further gives $\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{s}^l)] = 1$, and thus $\overline{m}[\mu_2(\mathbf{s}^l)] = \log\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{s}^l)] = 0$.

Next we turn to the term $\overline{m}[\mu_2(\mathbf{x}^l)]$. Again we use the definitions and notations from section B. We further denote $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \ldots, \lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^l = \boldsymbol{W}^l(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$. Using these notations, we get

$$\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{y}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^l, \boldsymbol{\alpha})_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^l_{\mathrm{c},:}\hat{\mathbf{r}}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right)^2\right]$$
$$= \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2\lambda_i. \tag{58}$$

Then due to $\boldsymbol{W}^l \overset{\theta^l}{\sim} \hat{\boldsymbol{W}}^l \overset{\theta^l}{\sim} \mathcal{N}(0, \sqrt{2\,/\,R_l}\boldsymbol{I})$;

$$\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)] = \frac{2}{R_l}\sum_i \lambda_i = \frac{2}{R_l}\operatorname{Tr}\boldsymbol{C}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$$
$$= 2\mu_2(\mathbf{x}^{l-1}). \tag{59}$$

where Eq. (59) follows from Corollary 3. Furthermore the symmetric propagation gives:

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[(\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}]^2 + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[(\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}]^2.$$
$$= \nu_{2,c}(\mathbf{y}^{l,+}) - \nu_{1,c}(\mathbf{y}^{l,+})^2 + \nu_{2,c}(\mathbf{y}^{l,-}) - \nu_{1,c}(\mathbf{y}^{l,-})^2$$
$$= \nu_{2,c}(\mathbf{y}^l) - \left(\nu_{1,c}(\mathbf{y}^{l,+})^2 + \nu_{1,c}(\mathbf{y}^{l,-})^2\right) \tag{60}$$

We have $\nu_{1,c}(\mathbf{y}^l) = \nu_{1,c}(\mathbf{y}^{l,+}) - \nu_{1,c}(\mathbf{y}^{l,-})$ and thus $\nu_{1,c}(\mathbf{y}^l)^2 = \nu_{1,c}(\mathbf{y}^{l,+})^2 + \nu_{1,c}(\mathbf{y}^{l,-})^2 - 2\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})$. We can then rewrite Eq. (60) as

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = \nu_{2,c}(\mathbf{y}^l) - \nu_{1,c}(\mathbf{y}^l)^2 - 2\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})$$
$$= \mu_{2,c}(\mathbf{y}^l) - 2\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-}) \tag{61}$$

Combining Eq. (59) and Eq. (61):

$$\begin{aligned}\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)+\mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l|A_{l-1}}[\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})],\\ 2\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l|A_{l-1}}[\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})], \qquad (62)\\ \mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= \mu_2(\mathbf{x}^{l-1})\left(1-\mathbb{E}_{\theta^l|A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right).\end{aligned}$$

where Eq. (62) is obtained by spherical symmetry of the propagation. We finally get

$$\begin{aligned}\mathbb{E}_{\theta^l|A_{l-1}}[\mu_2(\mathbf{x}^l)] &= \mathbb{E}_{\theta^l|A_{l-1}}\left[\mathbb{E}_{\mathrm{c}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)]\right] = \mathbb{E}_{\mathrm{c}}\left[\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)]\right]\\ &= \mu_2(\mathbf{x}^{l-1})\left(1-\mathbb{E}_{\mathrm{c},\theta^l|A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right),\\ \mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{x}^{l-1})] &= 1-\mathbb{E}_{\mathrm{c},\theta^l|A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right].\end{aligned}$$

Combined with $\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{s}^l)]=1$,

$$\begin{aligned}\delta\zeta^l \simeq \exp\left(\overline{m}_{\text{vanilla}}[\zeta^l]\right) &= \left(\frac{\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{s}^l)]}{\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{x}^l)]}\right)^{1/2}\\ &= \left(1-\mathbb{E}_{\mathrm{c},\theta^l|A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-1/2}.\end{aligned}$$

□

F.2 If ζ^l has drift larger than diffusion, then $\mu_2(\mathbf{x}^l)\,/\,\nu_2(\mathbf{x}^l)$ converges in probability to zero

We only need to slightly adapt the reasoning of section E.4. From theorem 1, we can write under A_l: $\log\nu_2(\mathbf{x}^l) = -lm_l + \sqrt{l}s_l + \log\nu_2(\mathbf{x}^0)$, with s_l centered and $m_{\min} \le m_l \le m_{\max}$, $v_{\min} \le \mathrm{Var}_{\Theta^l|A_l}[s_l] \le v_{\max}$. We can also write $\log\mu_2(\mathbf{s}^l) = -lm'_l + \sqrt{l}s'_l$, with s'_l centered and $m_{\min} \le m'_l \le m_{\max}$ and $v_{\min} \le \mathrm{Var}_{\Theta^l|A_l}[s'_l] \le v_{\max}$. We further suppose that there is an event D with $\mathbb{P}(D)>0$ under which ζ^l has *drift larger than diffusion*. To make it precise, this means that $\exists m > \frac{1}{2}(m_{\max}-m_{\min})$ such that for l large enough: $\log\delta\zeta^l \ge m$, and thus $\exists c\in\mathbb{R}$ such that for $\forall l$: $\log\zeta^l \ge c + lm$.

The ratio $\mu_2(\mathbf{x}^l)\,/\,\nu_2(\mathbf{x}^l)$ can be expressed as

$$\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} = \frac{\mu_2(\mathbf{x}^l)}{\mu_2(\mathbf{s}^l)\mu_2(\mathbf{x}^0)}\frac{\mu_2(\mathbf{s}^l)\mu_2(\mathbf{x}^0)}{\nu_2(\mathbf{x}^l)} = \frac{1}{(\zeta^l)^2}\frac{\mu_2(\mathbf{s}^l)\mu_2(\mathbf{x}^0)}{\nu_2(\mathbf{x}^l)},$$

which gives with logarithms,

$$\begin{aligned}\log\mu_2(\mathbf{x}^l) - \log\nu_2(\mathbf{x}^l) &= -2\log\zeta^l + \log\mu_2(\mathbf{s}^l) - \log\nu_2(\mathbf{x}^l) + \log\mu_2(\mathbf{x}^0)\\ &\le -2c - 2lm - lm_{\min} + \sqrt{l}s'_l + lm_{\max} - \sqrt{l}s_l - \log\nu_2(\mathbf{x}^0) + \log\mu_2(\mathbf{x}^0)\\ &\le C - lM + \sqrt{l}s_l,\end{aligned}$$

where we denoted $C = -2c - \log\nu_2(\mathbf{x}^0) + \log\mu_2(\mathbf{x}^0)$, $M = 2m + m_{\min} - m_{\max} > 0$ and $s_l = s'_l - s_l$. The variance of s_l is bounded as

$$\begin{aligned}\mathbb{E}_{\Theta^l|A_l}[s_l^2] &= \mathrm{Var}_{\Theta^l|A_l}[s_l] + \mathrm{Var}_{\Theta^l|A_l}[s'_l] - 2\mathbb{E}_{\Theta^l|A_l}[s_l s'_l]\\ &\le \mathrm{Var}_{\Theta^l|A_l}[s_l] + \mathrm{Var}_{\Theta^l|A_l}[s'_l] + 2\mathrm{Var}_{\Theta^l|A_l}[s_l]^{1/2}\,\mathrm{Var}_{\Theta^l|A_l}[s'_l]^{1/2} \le 4v_{\max}\\ \mathrm{Var}_{\Theta^l|A_l,D}[s_l] = \mathbb{E}_{\Theta^l|A_l,D}[s_l^2] &= \frac{1}{\mathbb{P}(D)}\mathbb{E}_{\Theta^l|A_l}[\mathbf{1}_D s_l^2] \le \frac{1}{\mathbb{P}(D)}4v_{\max}.\end{aligned}$$

Now for given ϵ:

$$\mathbb{P}_{\Theta^l|A_l,D}\left[\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon\right] = \mathbb{P}_{\Theta^l|A_l,D}\left[\log\mu_2(\mathbf{x}^l) - \log\mu_2(\mathbf{x}^l) > \log\epsilon\right]$$
$$\leq \mathbb{P}_{\Theta^l|A_l,D}\left[s_l > \frac{1}{\sqrt{l}}\big(lM + \log\epsilon - C\big)\right]$$

Chebyshev's inequality on the centered random variable s_l further gives

$$\mathbb{P}_{\Theta^l|A_l,D}\left[\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon\right] \leq \frac{l}{\big(lM + \log\epsilon - C\big)^2}\,\mathrm{Var}_{\Theta^l|A_l,D}[s_l]$$
$$\leq \frac{l}{\big(lM + \log\epsilon - C\big)^2}\,\frac{1}{\mathbb{P}(D)}4v_{\max} \sim \frac{1}{l}\,\frac{4v_{\max}}{M^2\mathbb{P}(D)},$$

proving that under D the ratio $\mu_2(\mathbf{x}^l)\,/\,\nu_2(\mathbf{x}^l)$ converges in probability to zero.

F.3 Limits of signal distribution

Proposition 12. *Suppose that*

$$\mathbb{P}_{\mathrm{c},\theta^l}\Big[\big\{\min\big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}),\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big) = 0\big\} \cap \big\{\max\big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}),\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big) > 0\big\}\Big] = 1.$$

Then $\mathbf{f}_{\mathrm{m}}(\mathbf{y}^l,\boldsymbol{\alpha})$ *concentrates on the semi-line generated by its average vector* $(\nu_{1,\mathrm{c}}(\mathbf{y}^l))_{1\leq \mathrm{c}\leq N_l}$.

Proof. Let us consider the feature map vectors $\mathbf{f}_{\mathrm{m}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})$ and receptive field vectors $\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})$. Due to the statistics-preserving property of Corollary 3, for each channel c and index $i_{\mathrm{c}} \in \mathcal{I}^l_{\mathrm{c}}$, $\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}} \overset{\mathbf{x},\boldsymbol{\alpha}}{\sim} RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}$ and thus $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{f}_{\mathrm{m}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_{\mathrm{c}}}]$.

Now let us reason by contradiction and suppose that there exists a direction $\boldsymbol{e}$ which is orthogonal to $\big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_{\mathrm{c}}}]\big)_{1\leq\, i_{\mathrm{c}}\leq R_l}$ and with non-zero variance $v > 0$. Consider the weight matrix $\boldsymbol{W}^l$ which projects on direction $\boldsymbol{e}$ for given channel c. For this direction, we have $\nu_{1,\mathrm{c}}(\mathbf{y}^l) = 0$, $\nu_{2,\mathrm{c}}(\mathbf{y}^l) > 0$. In turn, this implies $\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}) = \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})$ and $\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}) + \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) > 0$, which further gives $\min\big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}),\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big) > 0$. By continuity, $\min\big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}),\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big) > 0$ in a small neighborhood for the sampling of the weights $\boldsymbol{W}^l$, which contradicts the hypothesis. It follows that $\mathbf{r}_{\mathrm{f}}(\mathbf{x}^l,\boldsymbol{\alpha})$ concentrates on the direction generated by $\big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_{\mathrm{c}}}]\big)_{1\leq\, i_{\mathrm{c}}\leq R_l}$.

Now consider the weight matrix $\boldsymbol{W}^l$ which projects on the direction generated by $\big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_{\mathrm{c}}}]\big)_{1\leq\, i_{\mathrm{c}}\leq R_l}$ for given channel c. A similar argument gives $\min\big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}),\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big) = 0$, and thus that $\mathbf{f}_{\mathrm{m}}(\mathbf{y}^l,\boldsymbol{\alpha})_{\mathrm{c}}$ either concentrates in $\mathbb{R}^+$ or concentrates in $\mathbb{R}^-$. It follows that $\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})$ concentrates on the semi-line generated by $\big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_{\mathrm{c}}}]\big)_{1\leq\, i_{\mathrm{c}}\leq R_l}$. Under standard initialization, the image $\mathbf{f}_{\mathrm{m}}(\mathbf{y}^l,\boldsymbol{\alpha})$ of $\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})$ by the affine transform $\mathbf{f}_{\mathrm{m}}(\mathbf{y}^l,\boldsymbol{\alpha}) = \boldsymbol{W}^l\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha}) + \mathbf{b}^l = \boldsymbol{W}^l\mathbf{r}_{\mathrm{f}}(\mathbf{x}^{l-1},\boldsymbol{\alpha})$ thus concentrates on the semi-line generated by its average vector $(\nu_{1,\mathrm{c}}(\mathbf{y}^l))_{1\leq \mathrm{c}\leq N_l}$. □

G Normalized Sensitivity increments of batch-normalized feedforward nets

G.1 Proof of Theorem 3

Theorem 3. Normalized Sensitivity increments of batch-normalized feedforward nets. *The dominating term in the evolution of* ζ^l *can be decomposed as the sum of a term* $\overline{m}_{BN}[\zeta^l]$ *due to batch normalization and a term* $\overline{m}_{\phi}[\zeta^l]$ *due to the nonlinearity* ϕ*:*

$$\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right) = \left(\frac{\mu_2(\mathbf{s}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2} \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right]^{1/2},$$
$$\exp\left(\overline{m}_{\phi}\left[\zeta^l\right]\right) = \left(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}\left[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})\right]\right)^{-1/2},$$
$$\delta\zeta^l \simeq \exp\left(\overline{m}_{BN/FF}\left[\zeta^l\right]\right) = \exp\left(\overline{m}_{BN}\left[\zeta^l\right] + \overline{m}_{\phi}\left[\zeta^l\right]\right).$$

Proof. First let us decompose the dominating term as the product of two terms:

$$\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right) = \left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]}{\mu_2(\mathbf{s}^{l-1})}\right)^{1/2}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2},$$
$$\exp\left(\overline{m}_{\phi}\left[\zeta^l\right]\right) = \left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]}\right)^{1/2}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]}\right)^{-1/2},$$
$$\exp\left(\overline{m}_{BN/FF}\left[\zeta^l\right]\right) = \left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)]}{\mu_2(\mathbf{s}^{l-1})}\right)^{1/2}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)]}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2}$$
$$= \exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right)\exp\left(\overline{m}_{\phi}\left[\zeta^l\right]\right).$$

$\overline{m}_{BN}\left[\zeta^l\right]$ is a dominating term in the evolution of ζ^l from $(\mathbf{x}^{l-1}, \mathbf{s}^{l-1})$ to $(\mathbf{z}^l, \mathbf{u}^l)$, while $\overline{m}_{\phi}\left[\zeta^l\right]$ is a dominating term in the evolution of ζ^l from $(\mathbf{z}^l, \mathbf{u}^l)$ to $(\mathbf{x}^l, \mathbf{s}^l)$. These terms can be seen as the contribution to $\overline{m}_{BN/FF}\left[\zeta^l\right]$ of respectively batch normalization and ϕ. Now let us explicitate both terms.

Term $\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right)$**.** First we note that batch normalization directly gives $\mu_2(\mathbf{z}^l) = 1$ and thus $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)] = 1$. Now let us explicitate $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]$:

$$\forall \mathrm{c}: \ \mathbf{u}^l_{:,\mathrm{c}} = \frac{\mathbf{t}^l_{:,\mathrm{c}}}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)^{1/2}}, \qquad \forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{u}^l) = \frac{\mu_{2,\mathrm{c}}(\mathbf{t}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)},$$
$$\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)] = \mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{u}^l)] = \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right]$$

All together, we get

$$\exp\left(\overline{m}_{BN}\left[\zeta^l\right]\right) = \left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]}{\mu_2(\mathbf{s}^{l-1})}\right)^{1/2}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2}$$
$$= \left(\frac{\mu_2(\mathbf{s}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2}\mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right]^{1/2}$$

Term $\exp\left(\overline{m}_{\phi}\left[\zeta^l\right]\right)$**.** We consider the symmetric propagation for batch-normalized feedforward nets and again we denote $B_l = \left\{\forall k \leq l, \forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) = \mu_{2,\mathrm{c}}(\mathbf{u}^l)\right\} = \bigcap_{k=1}^{l}\left\{\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^k) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{s}}^k) = \mu_{2,\mathrm{c}}(\mathbf{u}^k)\right\}$ and B_l' the complementary event. By Eq. (33): $\mathbb{P}_{\Theta^l}(B_l) = 1$, $\mathbb{P}_{\Theta^l}(B_l') = 0$, and replicating the reasoning of Eq. (54) we deduce

$$\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)] + \mathbb{E}_{\theta^l}[\mu_2(\tilde{\mathbf{s}}^l)] = \mathbb{E}_{\theta^l|B_l}[\mu_2(\mathbf{s}^l)] + \mathbb{E}_{\theta^l|B_l}[\mu_2(\tilde{\mathbf{s}}^l)] = \mathbb{E}_{\theta^l|B_l}[\mu_2(\mathbf{u}^l)] = \mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)],$$
$$2\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)] = \mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)], \tag{63}$$

where Eq. (63) follows from spherical symmetry of the propagation. Now we turn to the symmetric propagation of the signal:

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[(\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}]^2 + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[(\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}]^2. \tag{64}$$
$$= \nu_{2,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{2,c}(\mathbf{z}^{l,-}) - \nu_{1,c}(\mathbf{z}^{l,-})^2$$
$$= \nu_{2,c}(\mathbf{z}^l) - \left(\nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2\right),$$

where Eq. (64) follows from Eq. (31). Due to the constraints imposed by batch normalization, $\nu_{1,c}(\mathbf{z}^l) = 0$ and $\nu_{2,c}(\mathbf{z}^l) = 1$, it follows that

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = 1 - \Big(\nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2\Big). \tag{65}$$

$$\nu_{1,c}(\mathbf{z}^l) = \nu_{1,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,-}) = 0,$$

$$\big(\nu_{1,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,-})\big)^2 = \nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2 - 2\nu_{1,c}(\mathbf{z}^{l,+})\nu_{1,c}(\mathbf{z}^{l,-}) = 0. \tag{66}$$

Using Eq. (65), Eq. (66) and the symmetry of the propagation,

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) = 1 - 2\nu_{1,c}(\mathbf{z}^{l,+})\nu_{1,c}(\mathbf{z}^{l,-}),$$

$$2\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)] = 1 - 2\mathbb{E}_{\mathrm{c},\theta^l}\Big[\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,-}\big)\Big]. \tag{67}$$

We finally combine Eq. (63) and Eq. (67):

$$\exp\left(\overline{m}_\phi\big[S^l\big]\right) = \left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]}\right)^{1/2}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]}\right)^{-1/2},$$

$$= \Big(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}\Big[\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,-}\big)\Big]\Big)^{-1/2}.$$

□

G.2 $\exp\left(\overline{m}_{BN}\big[\zeta^1\big]\right) > 1$ IN THE FIRST STEP OF THE PROPAGATION

Let us explicitate the second-order moment in channel c of $\mathbf{t}^1$:

$$\mu_{2,\mathrm{c}}(\mathbf{t}^1) = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\Big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^1,\boldsymbol{\alpha})_{\mathrm{c}}^2\Big] = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\Big[\mathbf{f}_{\mathrm{m}}(\mathbf{t}^1,\boldsymbol{\alpha})_{\mathrm{c}}^2\Big] = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^1_{\mathrm{c},:}\mathbf{r}_{\mathrm{f}}(\mathbf{s},\boldsymbol{\alpha})\big)^2\Big] \tag{68}$$

$$= \sum\nolimits_{i,j} \boldsymbol{W}^1_{\mathrm{c},i}\boldsymbol{W}^1_{\mathrm{c},j}\mathbb{E}_{\mathbf{s},\boldsymbol{\alpha}}[\mathbf{r}_{\mathrm{f}}(\mathbf{s},\boldsymbol{\alpha})_i\mathbf{r}_{\mathrm{f}}(\mathbf{s},\boldsymbol{\alpha})_j] = \sum\nolimits_i \big(\boldsymbol{W}^1_{\mathrm{c},i}\big)^2 = ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2. \tag{69}$$

where Eq. (68) follows from $\mathbf{t}^1$ being centered and Eq. (69) follows from the white noise property $\mathbb{E}_{\mathbf{s}}[\mathbf{s}_i\mathbf{s}_j] = \delta_{ij}$, which implies for any $\boldsymbol{\alpha}$ that $\mathbb{E}_{\mathbf{s}}[\mathbf{r}_{\mathrm{f}}(\mathbf{s},\boldsymbol{\alpha})_i\mathbf{r}_{\mathrm{f}}(\mathbf{s},\boldsymbol{\alpha})_j] = \delta_{ij}$ under periodic boundary conditions.

Now we turn to the second-order moment in channel c of $\mathbf{y}^1$. Denoting $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{R_1})$ and $(\lambda_1,\ldots,\lambda_{R_1})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}[\mathbf{r}_{\mathrm{f}}(\mathbf{x},\boldsymbol{\alpha})]$, and $\hat{\boldsymbol{W}}^1 = \boldsymbol{W}^1(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_{R_1})$, we get

$$\mu_{2,\mathrm{c}}(\mathbf{y}^1) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^1,\boldsymbol{\alpha})_{\mathrm{c}}^2\Big] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^1_{\mathrm{c},:}\hat{\mathbf{r}}_{\mathrm{f}}(\mathbf{x},\boldsymbol{\alpha})\big)^2\Big] = \sum\nolimits_i \big(\hat{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i$$

$$= ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i = \mu_{2,\mathrm{c}}(\mathbf{t}^1)\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i,$$

where we defined $\tilde{\boldsymbol{W}}^1$ such that $\forall c$: $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:} = \hat{\boldsymbol{W}}^1_{\mathrm{c},:} \,/\, ||\boldsymbol{W}^1_{\mathrm{c},:}||$ and we used Eq. (69). Under standard initialization, the distribution of $\boldsymbol{W}^1$ is spherically symmetric, implying that for all c the distribution of $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:}$ is uniform on the sphere of $\mathbb{R}^{R_1}$. In turn, this implies

$$\forall i:\ \mathbb{E}_{\theta^1}\Big[\big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\Big] = \frac{1}{R_1},$$

$$\forall c:\ \mathbb{E}_{\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big] = \frac{1}{R_1}\sum\nolimits_i \lambda_i, \qquad \mathbb{E}_{\mathrm{c},\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big] = \frac{1}{R_1}\sum\nolimits_i \lambda_i. \tag{70}$$

Finally we can write $\exp\left(\overline{m}_{BN}\big[\zeta^1\big]\right)$ as

$$\exp\left(\overline{m}_{BN}\big[\zeta^1\big]\right) = \left(\frac{\mu_2(\mathbf{s}^0)}{\mu_2(\mathbf{x}^0)}\right)^{-1/2}\mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^1)}{\mu_{2,\mathrm{c}}(\mathbf{y}^1)}\right]^{1/2}$$

$$= \Big(\frac{1}{R_1}\sum_i \lambda_i\Big)^{1/2}\mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{1}{\sum_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i}\right]^{1/2}, \tag{71}$$

$$\geq \Big(\frac{1}{R_1}\sum_i \lambda_i\Big)^{1/2}\left(\mathbb{E}_{\mathrm{c},\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big]^{-1}\right)^{1/2} = 1. \tag{72}$$

where Eq. (71) is obtained using $\mu_2(\mathbf{s}^0) = 1$ and $\mu_2(\mathbf{x}^0) = \frac{1}{R_1} \operatorname{Tr} \boldsymbol{C}[\mathbf{r}_{\mathrm{f}}(\mathbf{x}, \boldsymbol{\alpha})] = \frac{1}{R_1} \sum_i \lambda_i$ by Corollary 3 while Eq. (72) is obtained using the convexity of $x \to 1/x$ and Eq. (70).

H NORMALIZED SENSITIVITY INCREMENTS OF BATCH-NORMALIZED RESNETS

H.1 ADAPTATION OF THE PREVIOUS SETUP TO RESNETS

Before proceeding to the analysis, slight adaptations and forewords are necessary. Let us denote $\Theta^{l,h} = (\boldsymbol{\omega}^{1,1}, \boldsymbol{\beta}^{1,1}, \dots, \boldsymbol{\omega}^{1,H}, \boldsymbol{\beta}^{1,H}, \dots, \boldsymbol{\omega}^{l,1}, \boldsymbol{\beta}^{l,1}, \dots, \boldsymbol{\omega}^{l,h}, \boldsymbol{\beta}^{l,h})$ for the full set of parameters up to layer h in residual unit l and $\theta^{l,h} = \Theta^{l,h} | \Theta^{l,h-1}$ for the conditional set of parameters of layer h in residual unit l. We further denote $\Theta^l = \Theta^{l,H}$ and $\theta^l = \Theta^{l,H} | \Theta^{l-1,H}$ respectively the full and conditional sets of parameters at the granularity of the residual unit.

We now clarify to what extent Theorem 3 on batch-normalized feedforward nets still apply. First let us rewrite the propagation at layer $1 \leq h \leq H$ inside residual unit l with the pre-activation perspective:

$$\mathbf{z}^{l,h} = BN(\mathbf{y}^{l,h-1}), \qquad \mathbf{x}^{l,h} = \phi(\mathbf{z}^{l,h}), \qquad \mathbf{y}^{l,h} = \boldsymbol{\omega}^{l,h} * \mathbf{x}^{l,h} + \boldsymbol{\beta}^{l,h}, \tag{73}$$

$$\mathbf{u}^{l,h} = BN'(\mathbf{y}^{l,h-1}) \odot \mathbf{t}^{l,h-1}, \quad \mathbf{s}^{l,h} = \phi'(\mathbf{z}^{l,h}) \odot \mathbf{u}^{l,h}, \quad \mathbf{t}^{l,h} = \boldsymbol{\omega}^{l,h} * \mathbf{s}^{l,h}, \tag{74}$$

In the pre-activation perspective, each layer starts with $(\mathbf{y}^{l,h-1}, \mathbf{t}^{l,h-1})$ after the convolution and ends at $(\mathbf{y}^{l,h}, \mathbf{t}^{l,h})$ again after the convolution. The concrete effect is that in the first layer $h = 1$ of each residual unit l, batch normalization and ϕ are completely deterministic conditionally on Θ^{l-1}. This occurs again for $h \geq 2$ since batch normalization and ϕ are random conditionally on Θ^{l-1} but completely deterministic conditionally on $\Theta^{l,h-1}$. At even larger granularity, due to the aggregation $(\mathbf{y}^l, \mathbf{t}^l) = \sum_{k=0}^{l} (\mathbf{y}^{k,H}, \mathbf{t}^{k,H})$, the input signal $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ of each residual unit becomes more and more correlated between successive l and less and less dependent on the parameters θ^{l-k} of individual previous units.

Since batch normalization and the nonlinearity ϕ are the drivers of the evolution of ζ^l, this shift can be thought as attributing the parameters and thus the stochasticity of layer h to the layer $h - 1$. A possible strategy is thus to consider the evolution from $(\mathbf{x}^{l,h-1}, \mathbf{s}^{l,h-1})$ to $(\mathbf{x}^{l,h}, \mathbf{s}^{l,h})$ for layers $2 \leq h \leq H$. This strategy however does not work for the first layer $h = 1$, since the input signal $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ depends on the whole sequence of parameters Θ^{l-1} and not only on $\theta^{l-1,H}$. Another strategy consists in treating all moment-related quantities as deterministic instead of random. Symmetric propagation does not occur strictly in this case, but it still occurs in a *mean-field sense* when averaging over channel. So we expect Theorem 3 to remain valid.

H.2 LEMMA ON DOT-PRODUCT

Lemma 13. *For any random tensor* $\mathbf{u}$ *of* $\mathbb{R}^{n \times \cdots \times n \times N}$*:*

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0,$$

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]^2\right] \leq \frac{1}{N r_{\mathrm{eff}}(\mathbf{u})} \mu_2(\mathbf{u}) \mathbb{E}_{\theta^l}[\mu_2(\mathbf{y}^{l,H})],$$

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\mathbf{t},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0,$$

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\mathbf{t},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]^2\right] \leq \frac{1}{N r_{\mathrm{eff}}(\mathbf{u})} \mu_2(\mathbf{u}) \mathbb{E}_{\theta^l}[\mu_2(\mathbf{t}^{l,H})].$$

Proof. By spherical symmetry, moments of $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}$ and $-\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}} = \hat{\mathbf{f}}_{\mathrm{m}}(-\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}$ have the same distribution with respect to θ^l. It follows that

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = \mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \big(- \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\big)\right]\right],$$

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u}, \boldsymbol{\alpha})_{\mathrm{c}} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0.$$

Next we note that

$$\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big] = \frac{1}{N}\sum_{\mathrm{c}}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\Big],$$
$$= \frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big], \tag{75}$$

where $\langle\cdot\rangle$ denotes the standard dot product in $\mathbb{R}^N$. Let us denote $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_N)$ and $(\lambda_1,\ldots,\lambda_N)$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})]$. We further denote u_i the unit-variance components of $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})$ in the basis $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_N)$, and y_i the components of $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})$ in the basis $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_N)$. This gives

$$\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha}) = \sum_i \sqrt{\lambda_i}u_i\boldsymbol{e}_i, \quad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[u_i\big] = 0, \quad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[u_i^2\big] = 1,$$
$$\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha}) = \sum_i y_i\boldsymbol{e}_i.$$

Now we decompose each component y_i of $\mathbf{y}^{l,H}$ as:

$$\forall j : \alpha_{i,j} = \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}[y_iu_j], \qquad y_i = \sum_j \alpha_{i,j}u_j + z_i,$$

From this definition, we get

$$\forall j:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_iu_j\big] = 0, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_iu_i\big] = \alpha_{i,i}, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_i^2\big] = \sum_j \alpha_{i,j}^2 + \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_i^2\big],$$
$$\mu_2(\mathbf{y}^{l,H}) = \frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[\langle\mathbf{y}^{l,H},\mathbf{y}^{l,H}\rangle\big] = \frac{1}{N}\sum_i \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_i^2\big] = \frac{1}{N}\Big(\sum_{i,j}\alpha_{i,j}^2 + \sum_i \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_i^2\big]\Big). \tag{76}$$

The dot product can be computed in any orthogonal basis, so we use the basis $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_N)$:

$$\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big] = \sum_i \sqrt{\lambda_i}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}[y_iu_i] = \sum_i \sqrt{\lambda_i}\alpha_{i,i}.$$

Spherical symmetry implies that moments of $y_1\boldsymbol{e}_1 + \cdots + y_i\boldsymbol{e}_i + \cdots + y_N\boldsymbol{e}_N$ and $y_1\boldsymbol{e}_1 + \cdots - y_i\boldsymbol{e}_i + \cdots + y_N\boldsymbol{e}_N$ have the same distribution with respect to θ^l. It follows that

$$\forall j \neq i:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_iu_i\big]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_ju_j\big] \overset{\theta^l}{\sim} \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[-y_iu_i\big]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_ju_j\big],$$
$$\forall j \neq i:\ \alpha_{i,i}\alpha_{j,j} \overset{\theta^l}{\sim} (-\alpha_{i,i})\alpha_{j,j},$$
$$\forall j \neq i:\ \mathbb{E}_{\theta^l}\big[\alpha_{i,i}\alpha_{j,j}\big] = 0.$$

We deduce that

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big]^2\right] = \sum_i \lambda_i\mathbb{E}_{\theta^l}\big[\alpha_{i,i}^2\big].$$

Spherical symmetry also implies that the distribution of $\alpha_{i,j}$ with respect to θ^l is the same for all i. Denoting (β_j) such that $\forall i,j$: $\beta_j = \mathbb{E}_{\theta^l}\big[\alpha_{i,j}^2\big]$, we get combined with Eq. (76):

$$\mathbb{E}_{\theta^l}\big[\mu_2(\mathbf{y}^{l,H})\big] \geq \frac{1}{N}\Big(\sum_{i,j}\beta_j\Big) = \sum_i \beta_i,$$
$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big]^2\right] = \sum_i \lambda_i\beta_i \leq \lambda_{\max}\Big(\sum_i \beta_i\Big) \leq \lambda_{\max}\mathbb{E}_{\theta^l}\big[\mu_2(\mathbf{y}^{l,H})\big].$$

Finally combining with Eq. (75):

$$\begin{aligned}\mathbb{E}_{\theta^l}\Big[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big]^2\Big] &= \frac{1}{N^2}\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\big\langle\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\right]^2\right]\\ &\le \frac{1}{N^2}\lambda_{\max}\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right]\\ &\le \frac{1}{Nr_{\mathrm{eff}}(\mathbf{u})}\mu_2(\mathbf{u})\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right],\end{aligned}$$

where we used $\lambda_{\max}r_{\mathrm{eff}}(\mathbf{u}) = \sum_i \lambda_i = N\mu_2(\mathbf{u})$. The same analysis can be applied to $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{u},\boldsymbol{\alpha})$ and $\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l,H},\boldsymbol{\alpha})$.

□

Corollary 14. *Let us denote the dot products for $k,l \ge 0$ as*

$$Y_{k,l} = \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{k,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big], \quad T_{k,l} = \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{k,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big],$$

which combined with Eq. (18) implies

$$\begin{aligned}\sum_{k=0}^{l-1} Y_{k,l} &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big],\\ \sum_{k=0}^{l-1} T_{k,l} &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big].\end{aligned}$$

Then by spherical symmetry $\forall k,l$, $Y_{k,l}$ and $T_{k,l}$ are centered, and $\forall\{k,l\} \ne \{k',l'\}$: $\mathbb{E}_{\Theta^{\max(k,l,k',l')}}[Y_{k,l}Y_{k',l'}] = 0$, $\mathbb{E}_{\Theta^{\max(k,l,k',l')}}[T_{k,l}T_{k',l'}] = 0$. Furthermore:

$$\forall k < l:\ \mathbb{E}_{\Theta^l}\left[Y_{k,l}^2\right] \le \mathbb{E}_{\Theta^{l-1}}\left[\frac{1}{Nr_{\mathrm{eff}}(\mathbf{y}^{k,H})}\mu_2(\mathbf{y}^{k,H})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{y}^{l,H})]\right], \tag{77}$$

$$\forall k < l:\ \mathbb{E}_{\Theta^l}\left[T_{k,l}^2\right] \le \mathbb{E}_{\Theta^{l-1}}\left[\frac{1}{Nr_{\mathrm{eff}}(\mathbf{t}^{k,H})}\mu_2(\mathbf{t}^{k,H})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{t}^{l,H})]\right], \tag{78}$$

$$\forall l:\ \mathbb{E}_{\Theta^l}\left[\Big(\sum_{k=0}^{l-1} Y_{k,l}\Big)^2\right] \le \mathbb{E}_{\Theta^{l-1}}\left[\frac{1}{Nr_{\mathrm{eff}}(\mathbf{y}^{l-1})}\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{y}^{l,H})]\right], \tag{79}$$

$$\forall l:\ \mathbb{E}_{\Theta^l}\left[\Big(\sum_{k=0}^{l-1} T_{k,l}\Big)^2\right] \le \mathbb{E}_{\Theta^{l-1}}\left[\frac{1}{Nr_{\mathrm{eff}}(\mathbf{t}^{l-1})}\mu_2(\mathbf{t}^{l-1})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{t}^{l,H})]\right]. \tag{80}$$

H.3 Proof of Theorem 4

Theorem 4. Normalized Sensitivity increments of batch-normalized resnets. *Suppose that for all depth l we can bound the effective ranks $r_{\min} \lesssim r_{\mathrm{eff}}(\mathbf{y}^l), r_{\mathrm{eff}}(\mathbf{y}^{l,H}), r_{\mathrm{eff}}(\mathbf{t}^l), r_{\mathrm{eff}}(\mathbf{t}^{l,H})$, the second-order central moment $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and the feedforward increments inside residual units $\delta_{\min} \lesssim \delta\zeta^{l,h} \lesssim \delta_{\max}$. Denote $\rho_{\min} = \big((\delta_{\min})^{2H}\mu_{2,\min} - \mu_{2,\max}\big)\,/\,\mu_{2,\max}$ and $\rho_{\max} = \big((\delta_{\max})^{2H}\mu_{2,\max} - \mu_{2,\min}\big)\,/\,\mu_{2,\min}$, and further consider $\tau_{\min}$, $\tau_{\max}$ such that $\tau_{\min} < \rho_{\min}\,/\,2$ and $\tau_{\max} > \rho_{\max}\,/\,2$. Then:*

$$\forall l \ll Nr_{\min}:\quad \left(1+\frac{\rho_{\min}}{l+1}\right)^{1/2} \lesssim \delta\zeta^l \lesssim \left(1+\frac{\rho_{\max}}{l+1}\right)^{1/2}, \tag{81}$$

$$\forall l \gg 1:\quad \frac{1}{2}\rho_{\min}\log l \lesssim \log\zeta^l \lesssim \frac{1}{2}\rho_{\max}\log l, \tag{82}$$

$$\forall l \gg 1:\quad l^{\tau_{\min}} \lesssim \zeta^l \lesssim l^{\tau_{\max}}. \tag{83}$$

Proof. First we introduce the additional constants $\gamma_{\min} = (\delta_{\min})^{2H}$ and $\gamma_{\max} = (\delta_{\max})^{2H}$, so that we can write $\rho_{\min} = \left(\gamma_{\min}\mu_{2,\min} - \mu_{2,\max}\right) / \mu_{2,\max}$ and $\rho_{2,\max} = \left(\gamma_{\max}\mu_{2,\max} - \mu_{2,\min}\right) / \mu_{2,\min}$.

We also remind that the symbols $\lesssim$ in Theorem 4 denote inequalities up to small non-dominating terms. We write $a \lesssim b$ when $a(1+\delta_a) \leq b(1+\delta_b)$ with $|\delta_a| \ll 1$, $|\delta_b| \ll 1$ with high probability. We write $a \simeq b$ when $a(1+\delta_a) = b(1+\delta_b)$ with $|\delta_a| \ll 1$, $|\delta_b| \ll 1$ with high probability. Denoting $\wedge$ for the logical *and*, the following rules are easily verified:

$$\begin{gathered}(a \lesssim b) \wedge (a \gtrsim b) \iff (a \simeq b),\\ (a \lesssim b) \implies (-a \gtrsim -b),\\ (a \lesssim b) \implies (1/a \gtrsim 1/b),\\ (a \lesssim b) \wedge (c \lesssim d) \implies (ac \lesssim bd),\\ (a \lesssim b) \wedge (b \lesssim c) \implies a \lesssim c.\end{gathered}$$

Finally $(a \lesssim b) \wedge (c \lesssim d) \implies (a + c \lesssim b + d)$ under the condition that $\left\{|a+c| \ll |a| + |c|\right\}$ and $\left\{|b+d| \ll |b| + |d|\right\}$ are very small probability events. We keep these rules in mind in the course of this proof.

Proof of Eq. (81). Adopting the same notations as Corollary 14 and using $\mathbf{y}^l = \sum_{k=0}^{l} \mathbf{y}^{k,H}$ by Eq. (18), we get

$$\begin{aligned}\mu_2(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\Big[\sum\nolimits_{k,k'} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{k,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{y}^{k',H},\boldsymbol{\alpha})_{\mathrm{c}}\Big] = \sum\nolimits_{k,k'} Y_{k,k'}.\\ \mu_2(\mathbf{t}^l) &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\Big[\sum\nolimits_{k,k'} \hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{k,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\mathbf{f}}_{\mathrm{m}}(\mathbf{t}^{k',H},\boldsymbol{\alpha})_{\mathrm{c}}\Big] = \sum\nolimits_{k,k'} T_{k,k'}.\end{aligned}$$

Now using the hypotheses $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and $r_{\min} \lesssim r_{\mathrm{eff}}(\mathbf{y}^{l,H})$, combined with Eq. (77) from Corollary 14,

$$\begin{aligned}(l+1)\mu_{2,\min} + \sum\nolimits_{k\neq k'} Y_{k,k'} \lesssim \mu_2(\mathbf{y}^l) \lesssim (l+1)\mu_{2,\max} + \sum\nolimits_{k\neq k'} Y_{k,k'},\\ \mathbb{E}_{\Theta^l}\left[\left(\sum\nolimits_{k\neq k'} Y_{k,k'}\right)^2\right] \lesssim l(l+1)\frac{1}{N r_{\min}}\mu_{2,\max}^2,\\ \mathbb{E}_{\Theta^l}\left[\left|\sum\nolimits_{k\neq k'} Y_{k,k'}\right|\right] \lesssim (l+1)\frac{1}{\sqrt{N r_{\min}}}\mu_{2,\max}, \qquad (84)\end{aligned}$$

where Eq. (84) is obtained using Cauchy-Schwarz inequality. It follows that for large width $N \gg 1$, with high probability $\left|\sum_{k\neq k'} Y_{k,k'}\right| \ll (l+1)\mu_{2,\min}$ and $\left|\sum_{k\neq k'} Y_{k,k'}\right| \ll (l+1)\mu_{2,\max}$, which then gives

$$(l+1)\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^l) \lesssim (l+1)\mu_{2,\max}. \qquad (85)$$

We can write $(\zeta^l)^2$ as

$$\begin{aligned}(\zeta^l)^2 &= \frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^l)}{\mu_2(\mathbf{y}^l)} = \mu_2(\mathbf{y}^0)\frac{\mu_2(\mathbf{t}^{l-1}) + T_{l,l} + 2\sum_{k=0}^{l-1} T_{k,l}}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2\sum_{k=0}^{l-1} Y_{k,l}},\\ (\zeta^l)^2 &= (\zeta^{l-1})^2\,\frac{\mu_2(\mathbf{y}^{l-1}) + \frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}T_{l,l} + 2\frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}\sum_{k=0}^{l-1} T_{k,l}}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2\sum_{k=0}^{l-1} Y_{k,l}}. \qquad (86)\end{aligned}$$

Let us denote $\forall k \leq l$: $\tilde{T}_{k,l} = \frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}T_{k,l}$ and $Y_l = \sum_{k=0}^{l-1} Y_{k,l}$ and $\tilde{T}_l = \sum_{k=0}^{l-1} \tilde{T}_{k,l}$. Eq. (86) then gives

$$(\delta\zeta^l)^2 = \frac{(\zeta^l)^2}{(\zeta^{l-1})^2} = \frac{\mu_2(\mathbf{y}^{l-1}) + \tilde{T}_{l,l} + 2\tilde{T}_l}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2Y_l}. \qquad (87)$$

Furthermore we can bound $\tilde{T}_{l,l}$ as

$$\begin{gathered}\tilde{T}_{l,l} = \frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}\mu_2(\mathbf{t}^{l,H}) = \frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}(\zeta^{l-1})^2\prod\nolimits_h(\delta\zeta^{l,h})^2\frac{\mu_2(\mathbf{y}^{l,H})}{\mu_2(\mathbf{y}^0)},\\ \gamma_{\min}\mu_{2,\min} \lesssim \tilde{T}_{l,l} \lesssim \gamma_{\max}\mu_{2,\max}. \qquad (88)\end{gathered}$$

Combining Eq. (79) and Eq. (80) from Corollary 14 with Eq. (85), we get for the variance of the terms $\tilde{T}_l$ and Y_l:

$$\mathbb{E}_{\Theta^l}\left[Y_l^2\right] \lesssim \frac{1}{Nr_{\min}} l\mu_{2,\max}^2, \tag{89}$$

$$\mathbb{E}_{\Theta^l}\left[\left(\frac{\mu_2(\mathbf{y}^0)}{(\zeta^{l-1})^2}\sum_{k<l} T_{k,l}\right)^2\right] \lesssim \frac{1}{Nr_{\min}}\mathbb{E}_{\Theta^{l-1}}\left[\frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^{l-1})}{(\zeta^{l-1})^2}\mathbb{E}_{\theta^l}\left[\frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^{l,H})}{(\zeta^{l-1})^2}\right]\right],$$

$$\mathbb{E}_{\Theta^l}\left[\tilde{T}_l^2\right] \lesssim \frac{1}{Nr_{\min}}\mathbb{E}_{\Theta^{l-1}}\left[\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\tilde{T}_{l,l}\right]\right] \lesssim \gamma_{\max}\frac{1}{Nr_{\min}} l\mu_{2,\max}^2. \tag{90}$$

It follows that $\left|Y_l\right| \ll 1$ and $\left|\tilde{T}_l\right| \ll 1$ with high probability when $l \ll Nr_{\min}$. Combined with Eq. (87) and Eq. (88),

$$\frac{\mu_2(\mathbf{y}^{l-1}) + \gamma_{\min}\mu_{2,\min}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\max}} \lesssim (\delta\zeta^l)^2 \lesssim \frac{\mu_2(\mathbf{y}^{l-1}) + \gamma_{\max}\mu_{2,\max}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\min}},$$

$$1 + \frac{\gamma_{\min}\mu_{2,\min} - \mu_{2,\max}}{(l+1)\mu_{2,\max}} \lesssim (\delta\zeta^l)^2 \lesssim 1 + \frac{\gamma_{\max}\mu_{2,\max} - \mu_{2,\min}}{(l+1)\mu_{2,\min}}.$$

Finally for $l \ll Nr_{\min}$, we find that $\delta\zeta^l$ is bounded as

$$\left(1 + \frac{\rho_{\min}}{l+1}\right)^{1/2} \lesssim \delta\zeta^l \lesssim \left(1 + \frac{\rho_{\max}}{l+1}\right)^{1/2}.$$

□

Proof of Eq. (82). Using Eq. (89) and Eq. (90), we apply Cauchy-Schwarz inequality on Y_l and $\tilde{T}_l$:

$$\mathbb{E}_{\Theta^l}\left[\left|Y_l\right|\right] \lesssim \frac{1}{\sqrt{Nr_{\min}}}\sqrt{l}\mu_{2,\max},$$

$$\mathbb{E}_{\Theta^l}\left[\left|\tilde{T}_l\right|\right] \lesssim \sqrt{\gamma_{\max}}\frac{1}{\sqrt{Nr_{\min}}}\sqrt{l}\mu_{2,\max}.$$

Combined with Eq. (85), we deduce that for large width $\left|Y_l\right| \ll \mu_2(\mathbf{y}^l)$ and $\left|\tilde{T}_l\right| \ll \mu_2(\mathbf{y}^l)$ always hold with high probability. Due to $\mu_{2,\min} \lesssim Y_{l,l} \lesssim \mu_{2,\max}$ and to Eq. (88), we also have with high probability that $Y_{l,l} \ll \mu_2(\mathbf{y}^l)$ and $\tilde{T}_l \ll \mu_2(\mathbf{y}^l)$ when $l \gg 1$. Now let us take the logarithm in Eq. (87):

$$2\log\delta\zeta^l = \log\left(1 + \frac{1}{\mu_2(\mathbf{y}^{l-1})}\tilde{T}_{l,l} + \frac{2}{\mu_2(\mathbf{y}^{l-1})}\tilde{T}_l\right) - \log\left(1 + \frac{1}{\mu_2(\mathbf{y}^{l-1})}Y_{l,l} + \frac{2}{\mu_2(\mathbf{y}^{l-1})}Y_l\right).$$

When $l \gg 1$, all the terms added to 1 in the logarithm are $\ll 1$ with high probability. Therefore for $l > l_0 \gg 1$, we can write

$$2\log\delta\zeta^l \simeq \frac{1}{\mu_2(\mathbf{y}^{l-1})}\tilde{T}_{l,l} + \frac{2}{\mu_2(\mathbf{y}^{l-1})}\tilde{T}_l - \frac{1}{\mu_2(\mathbf{y}^{l-1})}Y_{l,l} - \frac{2}{\mu_2(\mathbf{y}^{l-1})}Y_l.$$

$$2\sum_{k=l_0+1}^{l}\log\delta\zeta^k \simeq \sum_{k=l_0+1}^{l}\frac{1}{\mu_2(\mathbf{y}^{k-1})}\left(\tilde{T}_{k,k} - Y_{k,k}\right) + \sum_{k=l_0+1}^{l}\frac{2}{\mu_2(\mathbf{y}^{k-1})}\left(\tilde{T}_k - Y_k\right).$$

Let us bound the first term:

$$\sum_{k=l_0+1}^{l}\frac{\gamma_{\min}\mu_{2,\min} - \mu_{2,\max}}{k\mu_{2,\max}} \lesssim \sum_{k=l_0+1}^{l}\frac{1}{\mu_2(\mathbf{y}^{k-1})}\left(\tilde{T}_{k,k} - Y_{k,k}\right) \lesssim \sum_{k=l_0+1}^{l}\frac{\gamma_{\max}\mu_{2,\max} - \mu_{2,\min}}{k\mu_{2,\min}},$$

$$\rho_{\min}\int_{l_0}^{l}\frac{1}{x}dx \lesssim \sum_{k=l_0+1}^{l}\frac{1}{\mu_2(\mathbf{y}^{k-1})}\left(\tilde{T}_{k,k} - Y_{k,k}\right) \lesssim \rho_{\max}\int_{l_0}^{l}\frac{1}{x}dx,$$

$$\rho_{\min}\log\left(\frac{l}{l_0}\right) \lesssim \sum_{k=l_0+1}^{l}\frac{1}{\mu_2(\mathbf{y}^{k-1})}\left(\tilde{T}_{k,k} - Y_{k,k}\right) \lesssim \rho_{\max}\log\left(\frac{l}{l_0}\right). \tag{91}$$

Now we consider the second term. By spherical symmetry, $Y_k \,/\, \mu_2(\mathbf{y}^{k-1})$ and $Y_{k'} \,/\, \mu_2(\mathbf{y}^{k'-1})$ are non-correlated for $k \neq k'$. So we get combined with Eq. (89) that for $l > l_0 \gg 1$,

$$\mathbb{E}_{\Theta^l}\left[\Big(\sum_{k=l_0+1}^{l} \frac{1}{\mu_2(\mathbf{y}^{k-1})} Y_k\Big)^2\right] = \sum_{k=l_0+1}^{l} \mathbb{E}_{\Theta^k}\left[\Big(\frac{1}{\mu_2(\mathbf{y}^{k-1})} Y_k\Big)^2\right],$$
$$\lesssim \frac{\mu_{2,\max}^2}{N r_{\min} \mu_{2,\min}^2} \sum_{k=l_0+1}^{l} \frac{k}{k^2} \lesssim \frac{\mu_{2,\max}^2}{N r_{\min} \mu_{2,\min}^2} \log\Big(\frac{l}{l_0}\Big).$$

A similar calculation for $\tilde{T}_k$ gives

$$\mathbb{E}_{\Theta^l}\left[\Big(\sum_{k=l_0+1}^{l} \frac{1}{\mu_2(\mathbf{y}^{k-1})} \tilde{T}_k\Big)^2\right] \lesssim \gamma_{\max} \frac{\mu_{2,\max}^2}{N r_{\min} \mu_{2,\min}^2} \log\Big(\frac{l}{l_0}\Big).$$

For large width and $l \gg 1$, these terms are very small with high probability compared to the term of Eq. (91). It follows that for $l > l_0 \gg 1$,

$$\rho_{\min} \log\Big(\frac{l}{l_0}\Big) \lesssim 2 \log\Big(\frac{\zeta^l}{\zeta^{l_0}}\Big) \lesssim \rho_{\max} \log\Big(\frac{l}{l_0}\Big),$$
$$\frac{1}{2}\rho_{\min} \log l \lesssim \log \zeta^l \lesssim \frac{1}{2}\rho_{\max} \log l.$$

□

Proof of Eq. (83). As a consequence Eq. (82), for $l \gg 1$ there exist δ, δ' with $|\delta| \ll 1$, $|\delta'| \ll 1$ with high probability and

$$(1+\delta)\frac{1}{2}\rho_{\min} \log l \leq \log \zeta^l \leq (1+\delta')\frac{1}{2}\rho_{\max} \log l,$$
$$\exp\Big(\frac{1}{2}(1+\delta)\rho_{\min} \log l\Big) \leq \zeta^l \leq \exp\Big(\frac{1}{2}(1+\delta')\rho_{\max} \log l\Big).$$

Now consider $\tau_{\min}$, $\tau_{\max}$ such that $\tau_{\min} < \frac{1}{2}\rho_{\min}$ and $\tau_{\max} > \frac{1}{2}\rho_{\max}$. Then

$$\frac{1}{2}(1+\delta)\rho_{\min} \log l = \big(\frac{1}{2}\rho_{\min} + \frac{1}{2}\rho_{\min}\delta - \tau_{\min}\big) \log l + \tau_{\min} \log l,$$
$$\frac{1}{2}(1+\delta')\rho_{\max} \log l = \big(\frac{1}{2}\rho_{\max} + \frac{1}{2}\rho_{\max}\delta' - \tau_{\max}\big) \log l + \tau_{\max} \log l.$$

The terms $\frac{1}{2}\rho_{\min} + \frac{1}{2}\rho_{\min}\delta - \tau_{\min}$ and $\frac{1}{2}\rho_{\max} + \frac{1}{2}\rho_{\max}\delta' - \tau_{\max}$ are respectively positive with high probability and negative with high probability. Therefore with high probability, $\exp\Big(\tau_{\min} \log l\Big) \leq \zeta^l \leq \exp\Big(\tau_{\max} \log l\Big)$ and thus for $l \gg 1$:

$$\exp\Big(\tau_{\min} \log l\Big) \lesssim \zeta^l \lesssim \exp\Big(\tau_{\max} \log l\Big),$$
$$l^{\tau_{\min}} \lesssim \zeta^l \lesssim l^{\tau_{\max}}.$$

□