Characterizing Well-behaved vs. Pathological Deep Neural Network Architectures

Antoine Labatie
antoine.labatie@centraliens.net

Abstract

We introduce a principled approach, requiring only mild assumptions, for the characterization of deep neural networks at initialization. Our approach applies both to fully-connected and convolutional networks and incorporates the commonly used techniques of batch normalization and skip-connections. Our key insight is to consider the evolution with depth of statistical moments of signal and sensitivity, thereby characterizing the well-behaved or pathological behaviour of input-output mappings encoded by different choices of architecture. We establish: (i) for feedforward networks with and without batch normalization, depth multiplicativity inevitably leads to ill-behaved moments and distributional pathologies; (ii) for residual networks, on the other hand, the mechanism of identity skip-connection induces power-law rather than exponential behaviour, leading to well-behaved moments and no distributional pathology.[1]

1 Introduction

Advances in the design of neural network architectures have more often come from the relentless race of practical applications rather than by principled approaches. Even after wide adoption, many common choices and rules of thumbs still await for theoretical validation. This is unfortunate since the emergence of a theory to fully characterize deep neural networks in terms of wanted and unwanted properties would enable to guide further improvements.

An important branch of research towards this theoretical characterization has focused on random networks at initialization, i.e. before any training has occured. The characterization of random networks offers valuable insights into the hypothesis space of input-output mappings reachable during training. In a Bayesian view, it also unveils the prior on the input-output mapping encoded by the choice of architecture (Neal, 1996a; Williams, 1997; Lee et al., 2017). Experimentally, well-behaved input-output mappings at initialization were extensively found to be predictive of model trainability and even post-training performance (Schoenholz et al., 2016; Balduzzi et al., 2017; Pennington et al., 2017; Yang & Schoenholz, 2017; Chen et al., 2018; Xiao et al., 2018).

Unfortunately, even this simplifying case of random networks is still challenging in various respects: (i) there is a complex interplay of different sources of randomness from input data and from model parameters; (ii) the finite number of units in each layer leads to additional complexity; (iii) convolutional neural networks – widely used in many applications – also lead to additional complexity. The difficulty (i) is typically circumvented by restricting input data to simplying cases: a one-dimensional input manifold (Poole et al., 2016; Raghu et al., 2017), two input data points (Schoenholz et al., 2016; Balduzzi et al., 2017; Yang & Schoenholz, 2017; Chen et al., 2018; Xiao et al., 2018), a batch of input data points (Anonymous, 2019). The difficulty (ii) is commonly circumvented by either considering the simplifying case of infinite width with the convergence of neural networks to Gaussian processes

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

Preprint. Work in progress.

(Neal, 1996a; Daniely et al., 2016; Lee et al., 2017; Matthews et al., 2018) or the simplifying case of typical activation patterns for ReLU networks (Balduzzi et al., 2017). Finally the difficulty (iii) is rarely tackled as most analyses only apply to fully-connected networks. To the best of our knowledge, all attempts have thus far been limited in their scope or simplifying assumptions.

In this paper, we introduce a principled approach to characterize random deep neural networks at initialization. Our approach does not require any of the usual simplifications of infinite width, gaussianity, typical activation patterns, restricted input data. It further applies both to fully-connected and convolutional networks, and incorporates batch normalization and skip-connections. Our key insight is to consider statistical moments of signal and sensitivity with respect to input data as random variables which depend on model parameters. By studying the evolution of these moments with depth, we characterize the well-behaved or pathological behaviour of input-output mappings encoded by different choices of architectures. Our findings span the topics of one-dimensional degeneracy, 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*.

Clean propagation. We consider a random tensorial input $\mathbf{x} \in \mathbb{R}^{n\times\cdots\times n\times N_0}$, spatially d-dimensional with extent n in all spatial dimensions and N_0 channels. This input $\mathbf{x}$ is fed into a d-dimensional convolutional neural network with periodic boundary conditions and fixed spatial extent equal to n.[1] For each layer l, 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, $\mathbf{x}^l, \mathbf{y}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ the tensors of post-activations and pre-activations. We further denote $\boldsymbol{\alpha}$ the spatial position, c the channel, and ϕ the activation function. Adopting the convention $\mathbf{x}^0 \equiv \mathbf{x}$, the propagation at each layer is given by

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

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 $\mathrm{d}\mathbf{x} \in \mathbb{R}^{n\times\cdots\times n\times N_0}$ with independent and identically distributed components such that $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i\,\mathrm{d}\mathbf{x}_j] = \sigma_{\mathrm{d}\mathbf{x}}^2\delta_{ij}$ ($\sigma_{\mathrm{d}\mathbf{x}} \ll 1$), with δ_{ij} the Kronecker delta. The noisy signal is propagated into the same neural network and we keep track of the noise corruption with the tensor $\mathrm{d}\mathbf{x}^l$ defined as $\mathrm{d}\mathbf{x}^0 \equiv \mathrm{d}\mathbf{x}$ and $\mathrm{d}\mathbf{x}^l \equiv \Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$, with $\mathbf{x}^l = \Phi_l(\mathbf{x})$ the neural network mapping from layer 0 to l. The simultaneous propagation of the *signal* $\mathbf{x}^l$ and the *noise* $\mathrm{d}\mathbf{x}^l$ is given by

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

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

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

To get rid of the dependence on $\sigma_{\mathrm{d}\mathbf{x}}$, the random *sensitivity tensor* is introduced as the rescaling $\mathbf{s}^0 \equiv \mathbf{s} \equiv \mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}}$ and $\mathbf{s}^l \equiv \mathrm{d}\mathbf{x}^l/\sigma_{\mathrm{d}\mathbf{x}}$. By 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 $\forall \mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\mathbf{s}}[\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}}] = 0$. The sensitivity tensor encodes *derivative information* while avoiding the burden of increased dimensionality (see calculations in Appendix D.3 and Appendix D.2). This will prove very useful.

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

Scope. We require two very mild assumptions: (i) $\mathbf{x}$ is not trivially 0 such that $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^2_{\boldsymbol{\alpha},\mathrm{c}}] > 0$;[2] (ii) the width N_l is bounded.

More importantly, we restrict to the – most commonly used – ReLU activation function $\phi(\cdot) = \max(\cdot, 0)$. Even though ϕ is not differentiable at 0, we still define $\mathrm{d}\mathbf{x}^l$ and $\mathbf{s}^l$ as the result of the simultaneous propagation of Eq. (1) and Eq. (2) with the convention $\phi'(0) \equiv 1/2$. Section 3 implicitly relies on the assumption that $\Phi_l(\mathbf{x})$ remains differentiable, implying $\mathrm{d}\mathbf{x}^l = \Phi_l(\mathbf{x} + \mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$, almost surely (a.s.) with respect to $\mathbf{x}$ (details and justification of this assumption in Appendix D.1).

Note that fully-connected neural networks are still included in our analysis as the subcase $n = 1$.

3 Input data randomness

We now turn our attention to the distributions with respect to input data of signal and sensitivity $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$. To outline the importance of these distributions, we may express by layer composition the output of an L-layer neural network as $\mathbf{x}^L = \Phi_{l,L}(\mathbf{x}^l)$, with $l < L$ and $\Phi_{l,L}$ the *upper neural network* mapping from layer l to L. It means that $\mathbf{x}^l$ and $\mathrm{d}\mathbf{x}^l$ can be seen as input signal and noise of this upper neural network. It also means that the distributions $P_{\mathbf{x}^l}(\mathbf{x}^l) = P_{\mathbf{x}}(\mathbf{x}^l)$ and $P_{\mathrm{d}\mathbf{x}^l}(\mathrm{d}\mathbf{x}^l) = P_{\mathbf{x},\mathrm{d}\mathbf{x}}(\mathrm{d}\mathbf{x}^l)$, or equivalently $P_{\mathbf{s}^l}(\mathbf{s}^l) = P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$, must remain well-behaved for this upper network to have a chance to accomplish its task successfully. We will return to this argument in Section 3.2 by detailing the penalizing effects of ill-behaved $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$.

3.1 Characterizing distributions with respect to input data

First we need to introduce the following statistical quantities to characterize $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$:

– The **feature map vector** and **centered feature map vector** associated with any random tensor $\mathbf{v}$,

$$\varphi(\mathbf{v}, \boldsymbol{\alpha}) \equiv \mathbf{v}_{\boldsymbol{\alpha},:}, \qquad \hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha}) \equiv \mathbf{v}_{\boldsymbol{\alpha},:} - \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha}}\big[\mathbf{v}_{\boldsymbol{\alpha},:}\big],$$

where $\boldsymbol{\alpha}$ is uniformly sampled in $\{1, \dots, n\}^d$ and the denotations $\varphi(\mathbf{v}, \boldsymbol{\alpha})$, $\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})$ remind the randomness of both $\mathbf{v}$ and $\boldsymbol{\alpha}$.

In a similar vein as batch normalization and Xiao et al. (2018), feature maps and centered feature maps at different spatial positions $\varphi(\mathbf{v}, \boldsymbol{\alpha})$ and $\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})$ are implicitly seen as sub-signals propagated together as part of tensorial meta-signals. Statistically, we work at the granularity of sub-signals and we treat equally the randomness from $\mathbf{v}$ and $\boldsymbol{\alpha}$.[3]

– The **non-central moment** and **central moment** of order p associated with any random tensor $\mathbf{v}$ per-channel and averaged over channels,

$$\nu_{p,\mathrm{c}}(\mathbf{v}) \equiv \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha}}\big[\varphi(\mathbf{v}, \boldsymbol{\alpha})^p_{\mathrm{c}}\big], \qquad \mu_{p,\mathrm{c}}(\mathbf{v}) \equiv \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha}}\big[\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})^p_{\mathrm{c}}\big],$$
$$\nu_{p}(\mathbf{v}) \equiv \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\big[\varphi(\mathbf{v}, \boldsymbol{\alpha})^p_{\mathrm{c}}\big], \qquad \mu_{p}(\mathbf{v}) \equiv \mathbb{E}_{\mathbf{v},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})^p_{\mathrm{c}}\big].$$

– The **effective rank** associated with any random tensor $\mathbf{v}$ (Vershynin, 2010),

$$r_{\mathrm{eff}}(\mathbf{v}) \equiv \frac{\mathrm{Tr}\, \boldsymbol{C}\big[\varphi(\mathbf{v}, \boldsymbol{\alpha})\big]}{||\boldsymbol{C}\big[\varphi(\mathbf{v}, \boldsymbol{\alpha})\big]||} = \frac{\sum_i \lambda_i}{\max_i \lambda_i},$$

where $\boldsymbol{C}\big[\,\cdot\,\big]$ denotes the covariance matrix, $||\cdot||$ the spectral norm, and (λ_i) the eigenvalues of $\boldsymbol{C}\big[\varphi(\mathbf{v}, \boldsymbol{\alpha})\big]$. The effective rank is always greater than 1, and it intuitively measures the number of effective directions which concentrate the variance of $\varphi(\mathbf{v}, \boldsymbol{\alpha})$.

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

[3] In the context e.g. of $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ note that propagation from layer 1 to l solely depends on the randomness of $\mathbf{x}$, the randomness of $\boldsymbol{\alpha}$ being only involved in the selection of the final position at layer l. For simplicity, $\boldsymbol{\alpha}$ will still be referred as input data.

– The **normalized sensitivity** – our key metric – derived from the moments of $\mathbf{x}^l$ and $\mathbf{s}^l$,

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

To better understand the definition of χ^l let us consider the classification task. Since the goal is to set apart different signals, the mean signal over $\mathbf{x}, \boldsymbol{\alpha}$ is uninformative. Centered feature maps $\hat{\varphi}(\mathbf{x}^l, \boldsymbol{\alpha})$ thus constitute the informative part of the signal and $\mu_2(\mathbf{x}^0)$, $\mu_2(\mathbf{x}^l)$ normalize by signal 'informative content' in Eq. (3). The normalized sensitivity χ^l exactly measures a root mean square sensitivity when neural network input and output signals $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^l$ are rescaled to have unit signal 'informative content': $\mu_2(\tilde{\mathbf{x}}) = 1$ and $\mu_2(\tilde{\mathbf{x}}^l) = 1$ (proof in Appendix D.2). If we denote the rescaled input-output mapping as $\tilde{\mathbf{x}}^l = \Psi_l(\tilde{\mathbf{x}})$, then this property is simply expressed as $\chi^l = \mathbb{E}_{\tilde{\mathbf{x}}}\left[\Psi_l'(\tilde{\mathbf{x}})^2\right]^{1/2}$ in the fully-connected case with 1-dimensional rescaled input and output signals $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{x}}^l$ ($N_0 = 1$, $N_l = 1$). We provide an illustration in Fig. 1.

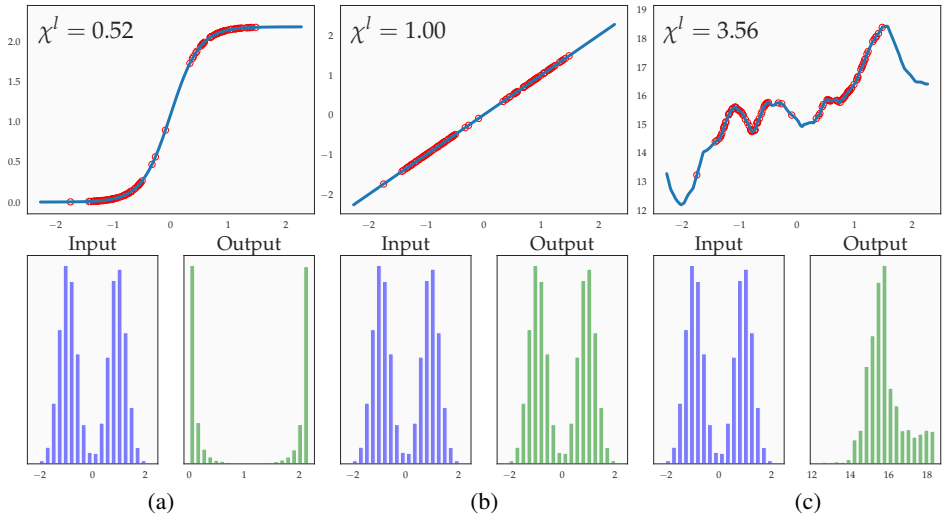

Figure 1: Illustration of the normalized sensitivity χ^l for fully-connected networks of l layers with 1-dimensional input and output rescaled signals $\tilde{\mathbf{x}}$ and $\mathbf{x}^l$. The distribution of the input $\tilde{\mathbf{x}}$ is a mixture of two Gaussians. We show the result of the propagation 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 $N_k = 100$ channels and ReLU activation for $1 \leq k < l$ and linear activation for $k = l$. *Top*: full input-output mapping (blue) and randomly sampled input-output data points (red circles). *Bottom*: histograms of inputs and outputs.

The normalized sensitivity χ^l is directly connected to the Jacobian norm in the fully-connected case and to sensitivity to signal perturbation in the general case, known in turn for being connected to generalization (Sokolic et al., 2017; Arora et al., 2018; Morcos et al., 2018; Novak et al., 2018; Philipp & Carbonell, 2018).[4] The notion of *sharpness* which quantifies sensitivity to weight perturbation was similarly shown to be connected to generalization (Hochreiter & Schmidhuber, 1997; Keskar et al., 2016; Smith & Le, 2017; Neyshabur et al., 2017). Sensitivity to signal perturbation and weight perturbation are tighly connected since the introduction of a noise $\mathrm{d}\boldsymbol{\omega}^l$ on the weights amounts to the introduction of a noise $\mathrm{d}\mathbf{y}^l = \mathrm{d}\boldsymbol{\omega}^l * \mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^l = \phi'(\mathbf{y}^l) \odot \mathrm{d}\mathbf{y}^l$ on the signal in Eq. (1) and Eq. (2).

[4]The coefficient defined in Philipp & Carbonell (2018) is equivalent to the normalized sensitivity χ^l in the fully-connected case. Section D.3 provides details on this equivalence and the reasons for our change of terminology.

3.2 Distributional pathologies

Using these statistical tools, we are able to characterize the distributional pathologies – with ill-behaved $P_{\mathbf{x}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathbf{s}}(\mathbf{s}^l)$ – that we will encounter:

– **Zero-dimensional signal**: $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$. To understand this pathology, let us consider the following mean vectors and rescaling of the signal:

$$\boldsymbol{\nu}^l \equiv (\nu_{1,\mathrm{c}}(\mathbf{x}^l))_{1\leq \mathrm{c}\leq N_l}, \qquad \tilde{\mathbf{x}}^l \equiv \frac{1}{||\boldsymbol{\nu}^l||_2}\mathbf{x}^l, \qquad \tilde{\boldsymbol{\nu}}^l \equiv (\nu_{1,\mathrm{c}}(\tilde{\mathbf{x}}^l))_{1\leq \mathrm{c}\leq N_l}.$$

Then $||\tilde{\boldsymbol{\nu}}^l||_2 = 1$ and the pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$ implies $\mu_2(\tilde{\mathbf{x}}^l) \xrightarrow{l\to\infty} 0$ (proof in Appendix D.4). It follows that $\varphi(\tilde{\mathbf{x}}^l, \boldsymbol{\alpha})$ becomes *point-like* concentrated at the point $\tilde{\boldsymbol{\nu}}^l$ of unit L^2 norm. In the limit of strict point-like concentration, the upper neural network from layer l to L is limited to *random guessing* since it 'sees' all inputs the same and cannot distinguish between them.

– **One-dimensional signal**: $r_{\text{eff}}(\mathbf{x}^l) \xrightarrow{l\to\infty} 1$. This pathology implies that $\varphi(\mathbf{x}^l, \boldsymbol{\alpha})$ has its variance which becomes concentrated in only one direction, and thus that $\varphi(\mathbf{x}^l, \boldsymbol{\alpha})$ becomes *line-like* concentrated. In the limit of strict line-like concentration, the upper neural network from layer l to L only 'sees' a *single feature* from $\mathbf{x}$.

– **Exploding sensitivity**: $\chi^l \xrightarrow{l\to\infty} \infty$. To understand this pathology, let us push further the view of noisy propagation with the *signal-to-noise ratio* SNR^l and the *noise factor* F^l:

$$SNR^l \equiv \frac{\mu_2(\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^l)}, \qquad F^l \equiv \frac{SNR^0}{SNR^l} = \frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)} = (\chi^l)^2, \tag{4}$$

where Eq. (4) follows from $\mu_2(\mathrm{d}\mathbf{x}^l) = \sigma^2_{\mathrm{d}\mathbf{x}}\mu_2(\mathbf{s}^l)$ and $\mu_2(\mathrm{d}\mathbf{x}^0) = \sigma^2_{\mathrm{d}\mathbf{x}}$. In logarithmic decibel scale $SNR^l_{\text{dB}} = SNR^0_{\text{dB}} - 20\log_{10}\chi^l$, indicating that χ^l measures how the neural network from layer 0 to l degrades ($\chi^l > 1$) or enhances ($\chi^l < 1$) the input signal-to-noise ratio. The pathology $\chi^l \xrightarrow{l\to\infty} \infty$ implies $F^l \xrightarrow{l\to\infty} \infty$ and $SNR^l \xrightarrow{l\to\infty} 0$, meaning that the noisy signal $\mathbf{x}^l + \mathrm{d}\mathbf{x}^l$ becomes completely random for any level of input noise $\mathrm{d}\mathbf{x}$.[5] The upper neural network from layer l to L is then limited to *random guessing*.

4 Model parameters randomness

We now introduce model parameters as the second source of randomness. We consider *random networks* at initialization, which we suppose 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.

Considering random networks at initialization is justified in two respects:

(i) From a *Bayesian perspective*, the random distribution on the weights encodes a prior distribution on input-output mappings for a given choice of architecture (Neal, 1996a; Williams, 1997). This prior distribution can be combined with the likelihood of model parameters given training data in order to obtain a posterior distribution on input-output mappings, which in turn enables Bayesian predictions (Lee et al., 2017).

(ii) If distributional pathologies arise at initilization, it is likely that training will be difficult. Indeed if the distributional pathology is severe, it becomes very difficult for the neural network to adjust its parameters and undo the pathology. As an illustration, consider the pathology of zero-dimensional signal: $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$. In this case, the upper neural network from layer l to L must adjust its bias parameters very precisely in order to center the signal and distinguish between different inputs $\mathbf{x}$. In support of this argument, several works have shown that deep neural networks are precisely trainable when they are not subject to the pathology

[5] In this case, Eq. (2) eventually does not hold since the noise $\mathrm{d}\mathbf{x}^l$ at layer l becomes non-infinetesimal even for infinitesimal input noise $\mathrm{d}\mathbf{x}$.

of different inputs $\mathbf{x}$ being strictly correlated at initialization, i.e. nearly undistinguishable (Schoenholz et al., 2016; Xiao et al., 2018).

From now on, our methodology is to consider all moment-related quantities, e.g. $\mu_p(\mathbf{x}^l)$, $\mu_p(\mathbf{s}^l)$, $\nu_p(\mathbf{x}^l)$, $\nu_p(\mathbf{s}^l)$, $r_{\text{eff}}(\mathbf{x}^l)$, $r_{\text{eff}}(\mathbf{s}^l)$, χ^l, as random variables depending on $(\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \dots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$. 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 $(\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \dots, \boldsymbol{\omega}^{l-1}, \boldsymbol{\beta}^{l-1})$ as given. We further denote the geometric increment $\delta\mu_2(\mathbf{x}^l) \equiv \mu_2(\mathbf{x}^l)/\mu_2(\mathbf{x}^{l-1})$.

Evolution with Depth. The evolution with depth of $\mu_2(\mathbf{x}^l)$ can be written as

$$\log \mu_2(\mathbf{x}^l) - \log \mu_2(\mathbf{x}^0) = \sum_{k\leq l} \underbrace{\log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]}_{\overline{m}[\mu_2(\mathbf{x}^k)]} + \sum_{k\leq l} \underbrace{\mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]}_{\underline{m}[\mu_2(\mathbf{x}^k)]} + \sum_{k\leq l} \underbrace{\log \delta\mu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)]}_{\underline{s}[\mu_2(\mathbf{x}^k)]}, \tag{5}$$

where Eq. (5) is obtained using $\log \mu_2(\mathbf{x}^l) - \log \mu_2(\mathbf{x}^0) = \sum_{k\leq l} \log \delta\mu_2(\mathbf{x}^k)$ and by expressing $\log \delta\mu_2(\mathbf{x}^k)$ with telescoping terms. Denoting $\underline{\delta}\mu_2(\mathbf{x}^k) \equiv \delta\mu_2(\mathbf{x}^k)/\mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]$ the multiplicatively centered increments, and using $\mathbb{E}_{\theta^k}[\log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]] = \log \mathbb{E}_{\theta^k}[\delta\mu_2(\mathbf{x}^k)]$ and $\mathbb{E}_{\theta^k}[\,\mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)]] = \mathbb{E}_{\theta^k}[\log \delta\mu_2(\mathbf{x}^k)]$, we obtain

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

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

$$\underline{s}[\mu_2(\mathbf{x}^k)] = \log \underline{\delta}\mu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \underline{\delta}\mu_2(\mathbf{x}^k)]. \tag{8}$$

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 $\mu_2(\mathbf{x}^k) = \frac{1}{N_k}\sum_c \mu_{2,c}(\mathbf{x}^k)$. As a consequence, for large N_k the relative increment $\underline{\delta}\mu_2(\mathbf{x}^k)$ has low expected deviation to 1, meaning with high probability that $|\log \underline{\delta}\mu_2(\mathbf{x}^k)| \ll 1$, $|\underline{m}[\mu_2(\mathbf{x}^k)]| \ll 1$, $|\underline{s}[\mu_2(\mathbf{x}^k)]| \ll 1$. 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 same analysis can be applied to other positive moments such as $\nu_2(\mathbf{x}^l)$, $\nu_2(\mathbf{s}^l)$, $\mu_2(\mathbf{s}^l)$. It can also be applied to the ratio $\mu_2(\mathbf{s}^l)/\mu_2(\mathbf{x}^l)$ and thus to χ^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. (6), (7) and (8) are extended to other central and non-central moments of signal and sensitivity, as well as χ^l with $\overline{m}[\chi^l] = \frac{1}{2}\big(\overline{m}[\mu_2(\mathbf{s}^l)] - \overline{m}[\mu_2(\mathbf{x}^l)]\big)$, $\underline{m}[\chi^l] = \frac{1}{2}\big(\underline{m}[\mu_2(\mathbf{s}^l)] - \underline{m}[\mu_2(\mathbf{x}^l)]\big)$, $\underline{s}[\chi^l] = \frac{1}{2}\big(\underline{s}[\mu_2(\mathbf{s}^l)] - \underline{s}[\mu_2(\mathbf{x}^l)]\big)$.

We introduce the notation $a \simeq b$ when $a(1+\epsilon_a) = b(1+\epsilon_b)$, with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$ with high probability. And the notation $a \lesssim b$ when $a(1+\epsilon_a) \leq b(1+\epsilon_b)$, with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$ with high probability. From now on, we assume that the *width is large*, implying $\exp(\underline{m}[\chi^l] + \underline{s}[\chi^l]) \simeq 1$ and $\delta\chi^l \simeq \exp(\overline{m}[\chi^l])$. We stress that this assumption is milder than the commonly used *mean-field* assumption of infinite width: $N_l \to \infty$. Indeed we still consider $\overline{m}[\chi^l]$ as random and $\mathbf{x}^l$ as possibly non-Gaussian, whereas mean-field would consider $\overline{m}[\chi^l]$ as non-random and $\mathbf{x}^l$ as Gaussian.

5 Vanilla Networks

We are fully equipped to characterize neural network architectures at initialization. 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) *There exist positive constants* m_{min}, m_{max}, v_{min}, $v_{\text{max}} > 0$, *random variables* $(m_l), (m'_l), (s_l), (s'_l)$ *and events* (A_l) *with probability* $\mathbb{P}_{\Theta^l}[A_l] \geq \prod_{k=1}^{l}(1 - 2^{-N_k+1})$ *such that under* A_l: s_l *and* 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. First we discuss the conditionality on A_l which is necessary to exclude the collapse $\nu_2(\mathbf{x}^l) = 0$ and $\mu_2(\mathbf{s}^l) = 0$ (with undefined $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$), occuring e.g. when all elements of $\boldsymbol{\omega}^l$ are strictly negative (Lu et al., 2018). The complementary event A^c has probability exponentially small in the width due to

$$-\mathbb{P}_{\Theta^l}[A_l^c] \simeq \log(1 - \mathbb{P}_{\Theta^l}[A_l^c]) = \log \mathbb{P}_{\Theta^l}[A_l] \geq \sum_{k=1}^{l} \log(1 - 2^{-N_k+1}) \simeq -\sum_{k=1}^{l} 2^{-N_k+1},$$

implying $\mathbb{P}_{\Theta^l}[A_l^c] \lesssim \sum_{k=1}^{l} 2^{-N_k+1}$. The conditionality on A_l thus has highly negligible effect.

Next we discuss the evolution of $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$ under A_l. The particularity of the initilization He et al. (2015) is to keep stable $\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},c}[(\mathbf{x}^l_{\boldsymbol{\alpha},c})^2] = \mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$ and $\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},c}[(\mathbf{s}^l_{\boldsymbol{\alpha},c})^2] = \mathbb{E}_{\Theta^l}[\mu_2(\mathbf{s}^l)]$ during propagation. To do so, it enforces $\mathbb{E}_{\theta^l}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1})$ and $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)] = \mu_2(\mathbf{s}^{l-1})$ such that $\overline{m}[\nu_2(\mathbf{x}^l)]$, $\overline{m}[\mu_2(\mathbf{s}^l)]$ vanish in Eq. (5) and $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathbf{s}^l)$ are subject to a *slow diffusion* with small negative drift terms: $\underline{m}[\nu_2(\mathbf{x}^l)] < 0$, $\underline{m}[\mu_2(\mathbf{s}^l)] < 0$, and small diffusion terms: $\underline{s}[\nu_2(\mathbf{x}^l)]$, $\underline{s}[\mu_2(\mathbf{s}^l)]$ (see Section E.3 for details).

As evidenced by Fig. 2a and Fig. 2b, the small negative drift and diffusion terms cause slowly decreasing negative expectation and increasing variance of $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathbf{s}^l)$. The decreasing negative expectation and increasing variance of $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathbf{s}^l)$ can be seen as opposing forces which compensate in order to keep stable $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$, $\mathbb{E}_{\Theta^l}[\mu_2(\mathbf{s}^l)]$ during propagation. Indeed Fig. 2a and Fig. 2b show that $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathbf{s}^l)$ are nearly Gaussian, implying that $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ are nearly lognormal. And the expectation of a lognormal variable $\exp(X)$ with $X \sim \mathcal{N}(\mu, \sigma)$ and $\mu < 0$ is equal to $\exp(\mu + \sigma^2/2)$. Note that the diffusion happens in log-space since layer composition amounts to a multiplicative random effect in real space. Also note that this is a finite-width effect since 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.

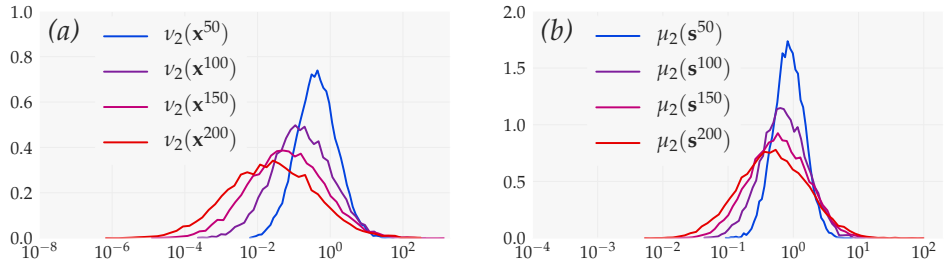

Figure 2: **Slowly diffusing moments of vanilla networks** with $N_l = 128$ and $L = 200$ layers. *(a)* Distributions of $\nu_2(\mathbf{x}^l)$ for $l = 50, 100, 150, 200$. *(b)* Distributions of $\mu_2(\mathbf{s}^l)$ for $l = 50, 100, 150, 200$. Both $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ show clear lognormality and are subject to a slow diffusion with small negative drift in log-space.

Theorem 2. Normalized Sensitivity increments of vanilla networks. (proof in Appendix E.4) *Denote* $\mathbf{y}^{l,+} = \max(\mathbf{y}^l, 0)$ *and* $\mathbf{y}^{l,-} = \max(-\mathbf{y}^l, 0)$. *The dominating term under* A_l *in the evolution of* χ^l *is*

$$\delta\chi^l \simeq \exp\left(\overline{m}_{\text{vanilla}}[\chi^l]\right) = \left(1 - \mathbb{E}_{c,\theta^l | A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-1/2}. \tag{9}$$

Discussion. An immediate consequence is that $\delta\chi^l \gtrsim 1$, i.e. that normalized sensitivity always increases with depth for random ReLU vanilla networks.

Now let us show that Theorem 1 and Theorem 2 only allow two possibilities of evolution, which both converge to distributional pathology:

(i) If sensitivity is exploding: $\chi^l \to \infty$, then $\mu_2(\mathbf{x}^l)/\mu_2(\mathbf{s}^l) \to 0$ by definition of χ^l. Given that $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) = \big(\mu_2(\mathbf{s}^l)/\nu_2(\mathbf{x}^l)\big)\big(\mu_2(\mathbf{x}^l)/\mu_2(\mathbf{s}^l)\big)$ and that both $\mu_2(\mathbf{s}^l)$ and $\nu_2(\mathbf{x}^l)$ are limited to the slow diffusion of Theorem 1, this qualitatively means that $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l)$ becomes $\ll 1$. Rigorously, if χ^l has an exponential drift stronger than diffusion and if we assume that $\nu_2(\mathbf{x}^l)$ and $\mu_2(\mathbf{s}^l)$ are lognormally-distributed as supported by Fig. 2, then $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \to 0$ a.s. (proof in Appendix E.5). Consequently, this case leads to double pathologies: *exploding sensitivity* and *zero-dimensional signal*.

(ii) Otherwise, geometric increments $\delta\chi^l$ are strongly limited and we may assume $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \to 1$. If we further assume that moments of the unit-variance rescaled signal $\tilde{\mathbf{x}}^l = \mathbf{x}^l/\sqrt{\mu_2(\mathbf{x}^l)}$ are bounded, then Theorem 2 implies the convergence to the pathology of *one-dimensional signal*: $r_{\text{eff}}(\mathbf{x}^l) \to 1$ (proof in Appendix E.6), as well as the convergence to neural network *pseudo-linearity* where each layer l becomes arbitrary well approximated by a linear function (proof in Appendix E.7)

Experimental verification. The evolution with depth of vanilla networks is shown in Fig. 3. From the possibilities (i) and (ii), it is case (ii) occuring with subexponential normalized sensitivity: $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \to 1$, the convergence to the pathology of one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$, and to neural network pseudo-linearity. A typical symptom of neural network pseudo-linearity is that signals 'cross' the ReLU on the same side. Our analysis offers a novel insight into this *coactivation* phenomenon, previously observed experimentally (Balduzzi et al., 2017).

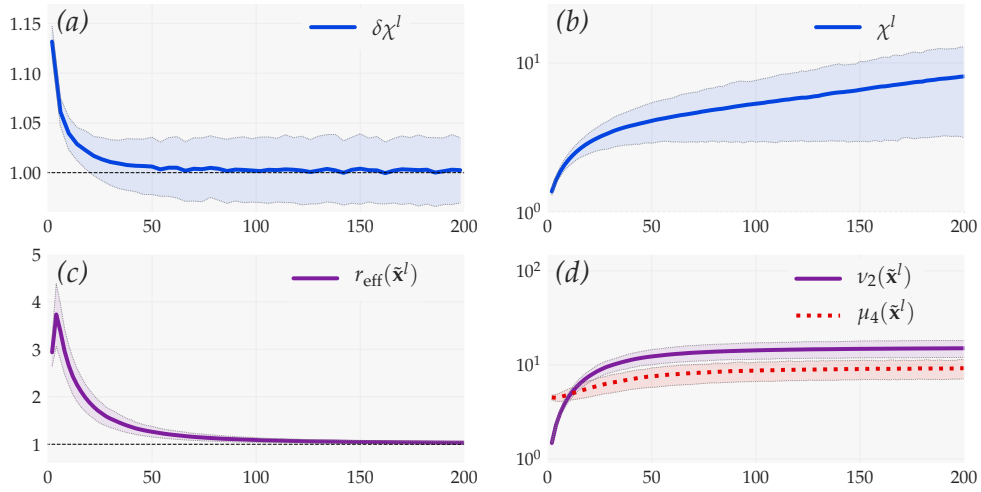

Figure 3: **Pathology of one-dimensional signal for vanilla networks** with $N_l = 384$ and $L = 200$ layers. *(a)* Geometric increments $\delta\chi^l$ with $\delta\chi^l \to 1$. *(b)* The growth of the normalized sensitivity is subexponential. *(c)* The effective rank $r_{\text{eff}}(\mathbf{x}^l)$ indicates one-dimensional pathology: $r_{\text{eff}}(\mathbf{x}^l) \to 1$. *(d)* Moments $\nu_2(\tilde{\mathbf{x}}^l)$ and $\mu_4(\tilde{\mathbf{x}}^l)$ of the unit-variance rescaled signal $\tilde{\mathbf{x}}^l = \mathbf{x}^l/\sqrt{\mu_2(\mathbf{x}^l)}$ remain well-behaved due to neural network pseudo-linearity.

6 Batch-normalized feedforward nets

Next we incorporate batch normalization (Ioffe & Szegedy, 2015). For simplicity, we only consider the *test mode* which consists in subtracting $\nu_{1,\text{c}}(\mathbf{y}^l)$ and dividing by $\sqrt{\mu_{2,\text{c}}(\mathbf{y}^l)}$ for each channel c in $\mathbf{y}^l$. The propagation is given by

$$\mathbf{y}^l = \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l, \quad \mathbf{z}^l = BN(\mathbf{y}^l), \qquad \mathbf{x}^l = \phi(\mathbf{z}^l), \tag{10}$$

$$\mathbf{t}^l = \boldsymbol{\omega}^l * \mathbf{s}^{l-1}, \qquad \mathbf{u}^l = BN'(\mathbf{y}^l) \odot \mathbf{t}^l, \quad \mathbf{s}^l = \phi'(\mathbf{z}^l) \odot \mathbf{u}^l, \tag{11}$$

where BN denotes batch normalization. Note that Eq. (10) and (11) 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 F.1) *The dominating term in the evolution of χ^l can be decomposed as the sum of a term $\overline{m}_{BN}[\chi^l]$ due to batch normalization and a term $\overline{m}_\phi[\chi^l]$ due to the nonlinearity ϕ:*

$$\exp\left(\overline{m}_{BN}\left[\chi^l\right]\right) = \left(\frac{\mu_2(\mathbf{s}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1/2} \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right]^{1/2}, \tag{12}$$

$$\exp\left(\overline{m}_{\phi}\left[\chi^l\right]\right) = \left(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}\left[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})\right]\right)^{-1/2}, \tag{13}$$

$$\delta\chi^l \simeq \exp\left(\overline{m}_{BN-FF}\left[\chi^l\right]\right) = \exp\left(\overline{m}_{BN}\left[\chi^l\right] + \overline{m}_{\phi}\left[\chi^l\right]\right). \tag{14}$$

Effect of batch normalization. The term of Eq. (12) corresponds to the evolution of χ^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 $\mathbf{x}^{l-1}$, while batch normalization can be seen as an alteration of the magnitude for each projection. Given that batch normalization uses $\sqrt{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}$ 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 χ^l due to the *noise factor equivalence* of Eq. (4).

Now let us look directly at the quantity inside the expectation in Eq. (12). 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 expectations $\mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)]/\mu_{2,\mathrm{c}}(\mathbf{x}^{l-1}) = \mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{t}^l)]/\mu_{2,\mathrm{c}}(\mathbf{s}^{l-1})$. On the other hand, the fluctuation of these increments depends on the fluctuation of 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. This effect of *conditioning* precisely depends on whether the effective ranks $r_{\text{eff}}(\mathbf{x}^{l-1})$ and $r_{\text{eff}}(\mathbf{s}^{l-1})$ are small compared to the ambient space dimension N_{l-1}.

If we assume that $\mathbf{s}^{l-1}$ is well-conditioned, then $\mu_{2,\mathrm{c}}(\mathbf{t}^l)$ has small relative deviation to its expectation $\mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{t}^l)]$ and this term can be treated as a constant. In turn, this implies by convexity of $x \mapsto 1/x$ that $\exp(\overline{m}_{BN}[\chi^l]) \gtrsim 1$. 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 $\mu_{2,\mathrm{c}}(\mathbf{y}^l)/\mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)]$ at the denominator and the impact of the convexity. Thus the smaller $r_{\text{eff}}(\mathbf{x}^{l-1})$ and the larger $\exp(\overline{m}_{BN}[\chi^l])$. This argument is strictly valid for the first step of the propagation where sensitivity has perfect conditioning, resulting in $\exp(\overline{m}_{BN}[\chi^1]) \geq 1$ (proof in Section F.2).

Effect of the nonlinearity ϕ. The term of Eq. (13) corresponds to the evolution of χ^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 $\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})/\mu_{2,\mathrm{c}}(\mathbf{z}^l)$ due to $\mu_{2,\mathrm{c}}(\mathbf{z}^l) = 1$ after batch normalization. We then find a very similar expression as Eq. (9) for vanilla networks.

Note that the separate attribution of $\exp(\overline{m}_{BN}[\chi^l])$ and $\exp(\overline{m}_{\phi}[\chi^l])$ to respectively batch normalization and the nonlinearity ϕ is still simplistic. In reality, both are working together since batch normalization maintains the exponential effect of ϕ by centering signal around 0. This is indeed necessary to ensure that, unlike $\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})$ in Eq. (9) for vanilla networks, $\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})$ in Eq. (13) remains non-vanishing for batch-normalized feedforward nets.

Experimental verification. In Fig. 4, we confirm experimentally the distributional pathology of *exploding sensitivity*: $\chi^l \to 0$. We also confirm that $\mathbf{s}^l$ is well-conditioned, while $\mathbf{x}^l$ becomes ill-conditioned. There is a clear inverse correlation between $r_{\text{eff}}(\mathbf{x}^l)$ and the effect of batch normalization on $\delta\chi^l$.

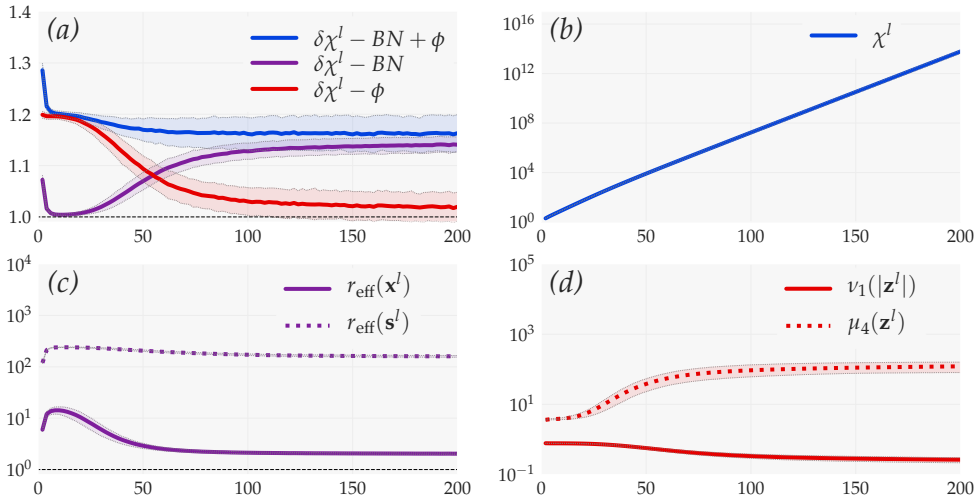

Figure 4: **Pathology of exploding sensitivity for batch-normalized feedforward nets** with $N_l = 384$ and $L = 200$ layers. *(a)* Geometric increments $\delta\chi^l$ decomposed as the evolution from $(\mathbf{x}^{l-1}, \mathbf{s}^{l-1})$ to $(\mathbf{z}^l, \mathbf{u}^l)$ corresponding to batch normalization, and the evolution from $(\mathbf{z}^l, \mathbf{u}^l)$ to $(\mathbf{x}^l, \mathbf{s}^l)$ corresponding to the nonlinearity ϕ. *(b)* The growth of χ^l indicates exploding sensitivity pathology: $\chi^l \to \infty$. *(c)* Effective ranks $r_{\text{eff}}(\mathbf{x}^l)$ and $r_{\text{eff}}(\mathbf{s}^l)$ of signal and sensitivity confirm that sensitivity is well-conditioned. There is a clear inverse correlation between $r_{\text{eff}}(\mathbf{x}^l)$ and the effect of batch normalization on $\delta\chi^l$. *(d)* $\mathbf{z}^l$ becomes ill-behaved with small $\nu_1(|\mathbf{z}^l|)$ and large $\mu_4(\mathbf{z}^l)$. This explains the decay of $\overline{m}_\phi[\chi^l]$ in Eq. (13), and thus of the effect of ϕ on $\delta\chi^l$.

Interestingly, the effect of the nonlinearity ϕ becomes subdominant as l grows. This is explained by the fact that $\mathbf{z}^l$ becomes heavy-tailed, with large $\mu_4(\mathbf{z}^l)$ and small $\nu_1(|\mathbf{z}^l|)$. Combined with $\nu_1(\mathbf{z}^{l,+}) \leq \nu_1(|\mathbf{z}^l|)$ and $\nu_1(\mathbf{z}^{l,-}) \leq \nu_1(|\mathbf{z}^l|)$, this explains the decay of $\overline{m}_\phi[\chi^l]$ in Eq. (13).

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 $N_l = N$,[6] and following He et al. (2016b) we adopt the perspective of *pre-activation units*. The propagation is given by

$$(\mathbf{y}^l, \mathbf{t}^l) = (\mathbf{y}^{l-1}, \mathbf{t}^{l-1}) + (\mathbf{y}^{l,H}, \mathbf{t}^{l,H}), \tag{15}$$

$$\underbrace{\begin{aligned} \mathbf{z}^{l,h} &= BN(\mathbf{y}^{l,h-1}), & \mathbf{x}^{l,h} &= \phi(\mathbf{z}^{l,h}), & \mathbf{y}^{l,h} &= \boldsymbol{\omega}^{l,h} * \mathbf{x}^{l,h} + \boldsymbol{\beta}^{l,h}, \\ \mathbf{u}^{l,h} &= BN'(\mathbf{y}^{l,h-1}) \odot \mathbf{t}^{l,h-1}, & \mathbf{s}^{l,h} &= \phi'(\mathbf{z}^{l,h}) \odot \mathbf{u}^{l,h}, & \mathbf{t}^{l,h} &= \boldsymbol{\omega}^{l,h} * \mathbf{s}^{l,h}. \end{aligned}}_{\text{for } 1 \leq h \leq H \text{ with } H \text{ the number of layers inside residual units, and with } (\mathbf{y}^{l,0}, \mathbf{t}^{l,0}) \equiv (\mathbf{y}^{l-1}, \mathbf{t}^{l-1})} \tag{16, 17}$$

If we adopt the additional convention $(\mathbf{y}^{0,H}, \mathbf{t}^{0,H}) \equiv (\mathbf{y}^0, \mathbf{t}^0)$, then Eq. (15) can be expanded as

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

For consistency reasons, we redefine the inputs of the propagation as $(\mathbf{y}^0, \mathbf{t}^0) \equiv (\mathbf{y}, \mathbf{t})$ and the normalized sensitivity and its increments as

$$\chi^l \equiv \left(\frac{\mu_2(\mathbf{t}^l)\, \mu_2(\mathbf{y}^0)}{\mu_2(\mathbf{y}^l)} \right)^{1/2}, \quad \delta\chi^l \equiv \frac{\chi^l}{\chi^{l-1}}, \quad \chi^{l,h} \equiv \left(\frac{\mu_2(\mathbf{t}^{l,h})\, \mu_2(\mathbf{y}^0)}{\mu_2(\mathbf{y}^{l,h})} \right)^{1/2}, \quad \delta\chi^{l,h} \equiv \frac{\chi^{l,h}}{\chi^{l,h-1}}.$$

[6] 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.

Theorem 4. Normalized Sensitivity increments of batch-normalized resnets. (proof in Appendix G.3) *Suppose that for all depth l we can bound the second-order central moments $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and the feedforward increments inside residual units $\delta_{\min} \lesssim \delta\chi^{l,h} \lesssim \delta_{\max}$.*[7] *Denote $\eta_{\min} = \big((\delta_{\min})^{2H}\mu_{2,\min} - \mu_{2,\max}\big)/\mu_{2,\max}$ and $\eta_{\max} = \big((\delta_{\max})^{2H}\mu_{2,\max} - \mu_{2,\min}\big)/\mu_{2,\min}$, as well as $\tau_{\min} = \eta_{\min}/2$ and $\tau_{\max} = \eta_{\max}/2$. Then there exist positive constants $C_{\min}$, $C_{\max} > 0$ such that*

$$\Big(1 + \frac{\eta_{\min}}{l+1}\Big)^{1/2} \lesssim \delta\chi^l \lesssim \Big(1 + \frac{\eta_{\max}}{l+1}\Big)^{1/2}, \tag{19}$$

$$C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}. \tag{20}$$

Discussion. The evolution in Eq. (19) remains exponential inside residual units since $\eta_{\min}$ and $\eta_{\max}$ 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 $(\mathbf{y}^{l,H}, \mathbf{t}^{l,H})$ into 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 set $\mu_{2,\min} = \mu_{2,\max}$ *and* we remove the dilution effect by replacing $l+1$ by 1, then Eq. (19) recovers the feedforward evolution with $(\delta_{\min})^H \lesssim \delta\chi^l \lesssim (\delta_{\max})^H$. The dilution is clearly visible as a side effect of 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 χ^l.

Given that $\log(1 + \frac{1}{x}) \simeq \frac{1}{x}$ for $x \gg 1$ and that $\int \frac{1}{x} dx = \log x$, the bounds in Eq. (20) are obtained by integration of the bounds in Eq. (19). A direct consequence of the dilution is thus the power-law evolution of χ^l in Eq. (20) instead of the exponential evolution for feedforward nets. Equivalently, when Eq. (20) is written as $C_{\min}\exp(\tau_{\min}\log l) \lesssim \chi^l \lesssim C_{\max}\exp(\tau_{\max}\log l)$, the evolution of χ^l for resnets is the same as the evolution of $\chi^{\log l}$ 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.

It is noteworthy that the benefits of adding skip connections (though in less sophisticated forms) were already known before the seminal works of He et al. (2016a) and He et al. (2016b). Both Duvenaud et al. (2014) and Neal (1996b) noted that the introduction of direct connections from the input to each layer alleviated distributional pathologies caused by layer composition. More recently, Philipp et al. (2017) also outlined the role of the dilution to avoid exploding gradients.

Experimental verification. The slow power-law evolution of χ^l is shown in Fig. 5. Notably, the exponent in the power-law fit of Fig. 5b is set to $\tau = \eta/2 = (\langle\delta\chi^{l,1}\rangle^{2H} - 1)/2$, with the feedforward increment averaged over the whole evolution $\langle\delta\chi^{l,1}\rangle$. This shows that in practice the evolution of χ^l is very well described by Eq. (20).

Contrary to batch-normalized feedforward nets, the signal remains well-behaved with: (i) many directions of signal variance preserved in $r_{\text{eff}}(\mathbf{x}^{l,1})$; (ii) close to Gaussian signal distribution as indicated by the moments $\nu_1(|\mathbf{z}^{l,1}|)$ and $\mu_4(\mathbf{z}^{l,1})$. No distributional pathology occurs.

8 Discussion and summary

This paper introduced a novel approach to characterize deep convolutional and fully-connected neural networks at initialization. The usual assumptions of infinite width, gaussianity, typical activation patterns, restricted input data were not required, and the commonly used techniques of batch normalization and skip connections were incorporated. The main scope restriction comes from our focus on the ReLU activation function. We leave for future work the application of our methodology to other activation functions – and potentially to recurrent neural networks.

It is interesting to look a posteriori at the commonly used assumption – not required in our analysis – that $\mathbf{x}^l$ is Gaussian with respect to input data $\mathbf{x}$, $\boldsymbol{\alpha}$ *and* model parameters Θ^l. Usually, the maximal

[7] The assumption $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ is very reasonable since batch normalization controls signal variance at the beginning of layer H: $\mu_2(\mathbf{z}^{l,H}) = 1$.

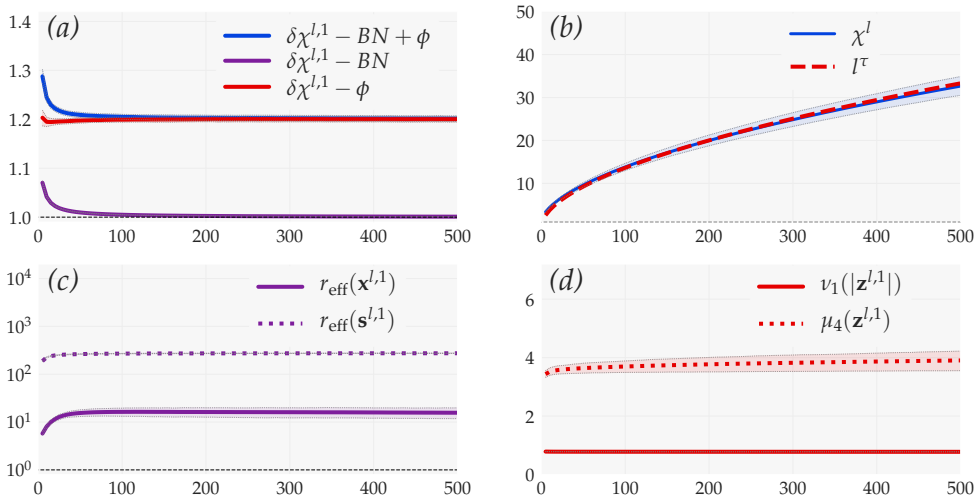

Figure 5: **Well-behaved evolution of batch-normalized resnets** with $N_l = 384$ and $L = 500$ residual units of $H = 2$ layers. *(a)* Geometric feedforward increments $\delta\chi^{l,1}$ decomposed as the product of a term corresponding to batch normalization and a term corresponding to ϕ. *(b)* χ^l has power-law evolution. *(c)* Effective ranks $r_{\text{eff}}(\mathbf{x}^{l,1})$ and $r_{\text{eff}}(\mathbf{s}^{l,1})$ of signal and sensitivity indicate that many directions of signal variance are preserved. *(d)* Moments $\nu_1(|\mathbf{z}^{l,1}|)$ and $\mu_4(\mathbf{z}^{l,1})$ indicate that $\mathbf{z}^{l,1}$ has close to Gaussian distribution.

depth L is considered constant and the width in the limit $N_l \to \infty$ for $l \leq L$. Our analysis has reversed this perspective by considering the depth in the limit $L \to \infty$ and the width N_l as bounded. Then the Gaussian assumption becomes invalid in two different contexts:

(i) In the case of vanilla networks, with e.g. an input $\varphi(\mathbf{x}, \boldsymbol{\alpha})$ reduced to a single point of $\mathbb{R}^{N_0}$ such that $\varphi(\mathbf{x}^l, \boldsymbol{\alpha})$ remains a single point of $\mathbb{R}^{N_l}$. Given the evolution of Fig. 3, the L^2 norm $||\varphi(\mathbf{x}^l, \boldsymbol{\alpha})||_2^2 = N_l \nu_2(\mathbf{x}^l)$ becomes lognormally-distributed with diverging variance with respect to Θ^l. This means that $\mathbf{x}^l$ becomes highly non-Gaussian.

(ii) In the case of batch-normalized feedforward nets, since batch normalization imposes $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}}] = \nu_{1,\text{c}}(\mathbf{z}^l) = 0$ and $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}[(\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}})^2] = \mu_{2,\text{c}}(\mathbf{z}^l) = 1$, it follows that $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\Theta^l}[\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}}] = 0$ and $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}})^2] = 1$. Given the evolution of Fig. 4, the fourth-order standardized moment with respect to $\mathbf{x}, \boldsymbol{\alpha}, \Theta^l$, i.e. the kurtosis, of the pre-activation signal $\mathbf{z}^l$ then becomes $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}})^4] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\text{c},\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\text{c}})^4] = \mathbb{E}_{\Theta^l}[\mu_4(\mathbf{z}^l)] \gg 1$. Again this means that $\mathbf{z}^l$ becomes highly non-Gaussian.

Similar observations were made in previous works. Duvenaud et al. (2014) showed that the composition of Gaussian processes, referred as deep Gaussian processes, leads to lognormal and eventually ill-behaved distributions. Matthews et al. (2018) found experimentally that convergence to gaussianity as $N_l \to \infty$ becomes slower with respect to N_l as l grows. This is explained by the fact that the *affine transform at each layer is an additive process* with respect to the width dimension, but *layer composition is a multiplicative process* with respect to the depth dimension. Qualitatively, the Central Limit Theorem implies that $\mathbf{x}^l$ becomes normally-distributed as $N_l \to \infty$ but log-normally distributed as $l \to \infty$.

Beside from this insight, our findings can be summarized as follows:

– In the case of vanilla networks, the initialization He et al. (2015) limits the evolution of second-order moments of signal and sensitivity. Combined with the limited growth of the normalized sensitivity χ^l, this results in the convergence to a distributional pathology where the neural network becomes linear and its signal shrunk to a single dimension: $r_{\text{eff}}(\mathbf{x}^l) \to 1$.

– In the case of batch-normalized feedforward nets, the pathology of exploding sensitivity with $\chi^l \to \infty$ has two origins: on the one hand batch normalization which upweights low-signal

pre-activation directions; on the other hand the nonlinearity ϕ. Beside from its direct effect of noise amplification, batch normalization also supports the exponential effect of ϕ on χ^l by keeping the signal centered around 0, where ϕ is the most nonlinear.

– Finally in the case of resnets, the normalized sensitivity χ^l 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 into the skip-connection path with ratio of signal variances decaying as $1/l$. 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 findings will open new perspectives in the statistical understanding of deep neural network architectures.

References

Anonymous. A mean field theory of batch normalization. In *Submitted to International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyMDXnCcF7. under review.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 254–263, 2018. URL http://proceedings.mlr.press/v80/arora18b.html.

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 *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 342–350, 2017. URL http://proceedings.mlr.press/v70/balduzzi17b.html.

Minmin Chen, Jeffrey Pennington, and Samuel S. Schoenholz. Dynamical isometry and a mean field theory of rnns: Gating enables signal propagation in recurrent neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 872–881, 2018. URL http://proceedings.mlr.press/v80/chen18i.html.

Marco Chiani, Davide Dardari, and Marvin K. Simon. New exponential bounds and approximations for the computation of error probability in fading channels. *IEEE Trans. Wireless Communications*, 2(4):840–845, 2003. doi: 10.1109/TWC.2003.814350. URL https://doi.org/10.1109/TWC.2003.814350.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 29*, pp. 2253–2261. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6427-toward-deeper-understanding-of-neural-networks-the-power-of-initialization-and-a-dual-view-on-expressivity.pdf.

Richard Durrett. *Probability: theory and examples*. Duxbury Press, second edition, 1996. ISBN 0-534-24318-5. URL https://services.math.duke.edu/~rtd/PTE/PTE4_1.pdf.

David Duvenaud, Oren Rippel, Ryan Adams, and Zoubin Ghahramani. Avoiding pathologies in very deep networks. In Samuel Kaski and Jukka Corander (eds.), *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 202–210, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL http://proceedings.mlr.press/v33/duvenaud14.html.

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)*, ICCV '15, pp. 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123. URL http://dx.doi.org/10.1109/ICCV.2015.123.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016a. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 630–645, 2016b. doi: 10.1007/978-3-319-46493-0_38. URL https://doi.org/10.1007/978-3-319-46493-0_38.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Comput.*, 9(1):1–42, January 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.1. URL http://dx.doi.org/10.1162/neco.1997.9.1.1.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 448–456, 2015. URL http://jmlr.org/proceedings/papers/v37/ioffe15.html.

Nitish S. Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping T. P. Tang. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. *ArXiv e-prints*, September 2016. URL http://adsabs.harvard.edu/abs/2016arXiv160904836S.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. *ArXiv e-prints*, October 2017. URL http://adsabs.harvard.edu/abs/2017arXiv171100165L.

Lu Lu, Yanhui Su, and George E. Karniadakis. Collapse of deep and narrow neural nets. *CoRR*, abs/1808.04947, 2018. URL http://arxiv.org/abs/1808.04947.

Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. Gaussian Process Behaviour in Wide Deep Neural Networks. *ArXiv e-prints*, April 2018. URL http://adsabs.harvard.edu/abs/2018arXiv180411271M.

Ari S. Morcos, David G.T. Barrett, Neil C. Rabinowitz, and Matthew Botvinick. On the importance of single directions for generalization. *ArXiv e-prints*, March 2018. URL http://adsabs.harvard.edu/abs/2018arXiv180306959M.

Radford M. Neal. *Priors for Infinite Networks*, pp. 29–53. Springer New York, New York, NY, 1996a. ISBN 978-1-4612-0745-0. doi: 10.1007/978-1-4612-0745-0_2. URL https://doi.org/10.1007/978-1-4612-0745-0_2.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer-Verlag, Berlin, Heidelberg, 1996b. ISBN 0387947248.

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. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7176-exploring-generalization-in-deep-learning.pdf.

Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. volume abs/1802.08760, 2018. URL http://arxiv.org/abs/1802.08760.

Jeffrey Pennington, Samuel S. Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 4788–4798, 2017. URL http://papers.nips.cc/paper/7064-resurrecting-the-sigmoid-in-deep-learning-through-dynamical//-isometry-theory-and-practice.

George Philipp and Jaime G. Carbonell. The nonlinearity coefficient - predicting overfitting in deep neural networks. *CoRR*, abs/1806.00179, 2018. URL http://arxiv.org/abs/1806.00179.

George Philipp, Dawn Song, and Jaime G. Carbonell. Gradients explode - deep networks are shallow - resnet explained. *CoRR*, abs/1712.05577, 2017. URL http://arxiv.org/abs/1712.05577.

Ben Poole, Subhaneil Lahiri, Maithreyi Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3360–3368, 2016. URL http://papers.nips.cc/paper/6322-exponential-expressivity-in-deep-neural-networks-through-transient-chaos.

Maithra Raghu, Ben Poole, Jon Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/raghu17a.html.

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

Samuel L. Smith and Quoc V. Le. A bayesian perspective on generalization and stochastic gradient descent. *CoRR*, abs/1710.06451, 2017. URL http://arxiv.org/abs/1710.06451.

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. doi: 10.1109/TSP.2017.2708039. URL https://doi.org/10.1109/TSP.2017.2708039.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. volume abs/1011.3027, 2010. URL http://arxiv.org/abs/1011.3027.

Christopher K. I. Williams. Computing with infinite networks. In M. C. Mozer, M. I. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems 9*, pp. 295–301. MIT Press, 1997. URL http://papers.nips.cc/paper/1197-computing-with-infinite-networks.pdf.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

Greg Yang and Samuel S. Schoenholz. Mean field residual networks: On the edge of chaos. *CoRR*, abs/1712.08969, 2017. URL http://arxiv.org/abs/1712.08969.

A Details of the experiments

All four experiments of Fig. 2, 3, 4, 5 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.

In Fig. 2, we considered the width $N_l = 128$ and the total depth $L = 200$. For $10\,000$ random realizations of He et al. (2015), 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 $10\,000$ realizations. The limited width – slightly less than standard values – had two purposes: (i) outlining the diffusion process, stronger for smaller width; (ii) limiting computation time in order to gather more realizations.

In Fig. 3, 4, 5, we increased the width to a more realistic value $N_l = 384$ and considered $1\,000$ realization with single batches of 64 randomly sampled images. For each realization we computed the empirical expectation over all realizations as well as 1σ intervals shown as shaded areas. Decreasing the number of inputs for each realization enabled to reduce computation time, mostly penalized by the computation of $r_{\text{eff}}(\mathbf{x}^l)$ and $r_{\text{eff}}(\mathbf{s}^l)$ involving an eigenvalue decomposition. We found that this reduction had very little impact for vanilla networks. As for batch-normalized feedforward nets and batch-normalized resnets, it enabled to match the usual setup of batch normalization in training mode.

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 $(\boldsymbol{\omega}^l * \mathbf{v})_{\boldsymbol{\alpha},:}$ at position $\boldsymbol{\alpha} \in \{1,\dots,n\}^d$ is obtained by the application of the convolution kernel $\boldsymbol{\omega}^l$ 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 $(\boldsymbol{\omega}^l * \mathbf{v})$ at spatial position $\boldsymbol{\alpha}$.

The *receptive field mapping* $RF(\cdot)$ associates an input $\mathbf{v}$ from layer $l-1$ to the tensor $RF(\mathbf{v}) \in \mathbb{R}^{n\times\cdots\times n\times K_l^d N_{l-1}}$, defined such that $RF(\mathbf{v})_{\boldsymbol{\alpha},:}$ is the reshaped vectorial form of the receptive field of $(\boldsymbol{\omega}^l * \mathbf{v})$ 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* ρ and *centered receptive field vector* $\hat{\rho}$ associated with an input $\mathbf{v}$ from layer l are random vectors which depend on $\mathbf{v}$, $\boldsymbol{\alpha}$ such that

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

where we kept the same denotation $\mathbf{v}$ for the variable in the expectation, as a slight abuse of notation. Again ρ and $\hat{\rho}$ 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: $RF(\mathbf{v})_{\boldsymbol{\alpha},i_{\mathrm{c}}} \sim_{\mathbf{v},\boldsymbol{\alpha}} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$.

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 $\mathbf{y}^l_{\boldsymbol{\alpha},:}$ can be written as

$$\mathbf{y}^l_{\boldsymbol{\alpha},:} = \boldsymbol{W}^l RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} + \mathbf{b}^l, \tag{21}$$

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. (21) 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

$$\mathbf{y}^l = \boldsymbol{W}^l RF(\mathbf{x}^{l-1}) + \mathbf{b}^l, \quad \mathbf{x}^l = \phi(\mathbf{y}^l), \tag{22}$$

$$\mathbf{t}^l = \boldsymbol{W}^l RF(\mathbf{s}^{l-1}), \qquad \mathbf{s}^l = \phi'(\mathbf{y}^l) \odot \mathbf{t}^l. \tag{23}$$

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

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

Gramian and Covariance. The *Gramian matrix* $\boldsymbol{G}$ and *covariance matrix* $\boldsymbol{C}$ associated with 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 gramian matrix and covariance matrix of feature map vectors and receptive field vectors are related by

$$\boldsymbol{G}[\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})] = \boldsymbol{C}[\hat{\varphi}(\mathbf{v}, \boldsymbol{\alpha})] = \boldsymbol{C}[\varphi(\mathbf{v}, \boldsymbol{\alpha})], \qquad \boldsymbol{G}[\hat{\rho}(\mathbf{v}, \boldsymbol{\alpha})] = \boldsymbol{C}[\hat{\rho}(\mathbf{v}, \boldsymbol{\alpha})] = \boldsymbol{C}[\rho(\mathbf{v}, \boldsymbol{\alpha})].$$

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:

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

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

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{y}^l). \tag{24}$$

Now we consider the second-order moments of the sensitivity tensor. We have the following identity:

$$\begin{aligned} (\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\bar{\mathbf{s}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 &= (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\bar{\mathbf{t}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \phi'(\bar{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 [\phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + \phi'(\bar{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2] \\ &= (\mathbf{t}^l_{\boldsymbol{\alpha},\mathrm{c}})^2, \end{aligned} \tag{25}$$

where Eq. (25) is obtained using $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} = -\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and the convention $\phi'(0) \equiv 1/2$. Since $\mathbf{s}^l$, $\bar{\mathbf{s}}^l$, $\mathbf{t}^l$ are centered, it follows that

$$\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{s}}^l) = \nu_{2,\mathrm{c}}(\mathbf{s}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{s}}^l) = \nu_{2,\mathrm{c}}(\mathbf{t}^l) = \mu_{2,\mathrm{c}}(\mathbf{t}^l). \tag{26}$$

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

$$\bar{\mathbf{y}}^l = -\boldsymbol{W}^l RF(\mathbf{x}^{l-1}) - \mathbf{b}^l, \quad \bar{\mathbf{z}}^l = BN(\bar{\mathbf{y}}^l), \qquad \bar{\mathbf{x}}^l = \phi(\bar{\mathbf{z}}^l), \tag{27}$$

$$\bar{\mathbf{t}}^l = -\boldsymbol{W}^l RF(\mathbf{s}^{l-1}), \qquad \bar{\mathbf{u}}^l = BN'(\bar{\mathbf{y}}^l) \odot \bar{\mathbf{t}}^l, \quad \bar{\mathbf{s}}^l = \phi'(\bar{\mathbf{z}}^l) \odot \bar{\mathbf{u}}^l. \tag{28}$$

BN in Eq. (27) and (28) uses the statistics of $\bar{\mathbf{y}}^l$ such that, under standard initialization, tensor moments have the *same distribution with respect to θ^l for both propagations*. We then simply have

$$\bar{\mathbf{z}}^l = -\mathbf{z}^l, \qquad \bar{\mathbf{u}}^l = -\mathbf{u}^l. \tag{29}$$

The exact same analysis as before gives

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{z}^l), \tag{30}$$

$$\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{s}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{s}}^l) = \mu_{2,\mathrm{c}}(\mathbf{u}^l). \tag{31}$$

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 $i_\mathrm{c} \in \mathcal{I}_\mathrm{c}^l$, 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 $\boldsymbol{\kappa} \in \{1,\dots,K_l\}^d$. 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 $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} = \mathbf{v}_{f(\boldsymbol{\alpha}),\mathrm{c}} \sim_{\boldsymbol{\alpha}} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ when $\boldsymbol{\alpha}$ is considered as random and $\mathbf{v}$ as given. In turn, this implies that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \sim_{\mathbf{v},\boldsymbol{\alpha}} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$, when both $\mathbf{v}$ and $\boldsymbol{\alpha}$ are considered as random.

□

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 $i_\mathrm{c} \in \mathcal{I}_\mathrm{c}^l$, we have $\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \sim_{\mathbf{x},\boldsymbol{\alpha}} \varphi(\mathbf{x}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$ and $\rho(\mathbf{s}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \sim_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}} \varphi(\mathbf{s}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$. Since the cardinality $|\mathcal{I}_\mathrm{c}^l| = K_l^d$ is the same for all channels* c*, it follows that*

$$\begin{aligned}
\nu_2(\mathbf{x}^{l-1}) &= \frac{1}{N_{l-1}} \operatorname{Tr} \boldsymbol{G}[\varphi(\mathbf{x}^{l-1},\boldsymbol{\alpha})] = \frac{1}{R_l} \operatorname{Tr} \boldsymbol{G}[\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha})],\\
\mu_2(\mathbf{x}^{l-1}) &= \frac{1}{N_{l-1}} \operatorname{Tr} \boldsymbol{C}[\varphi(\mathbf{x}^{l-1},\boldsymbol{\alpha})] = \frac{1}{R_l} \operatorname{Tr} \boldsymbol{C}[\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha})].\\
\nu_2(\mathbf{s}^{l-1}) = \mu_2(\mathbf{s}^{l-1}) &= \frac{1}{N_{l-1}} \operatorname{Tr} \boldsymbol{C}[\varphi(\mathbf{s}^{l-1},\boldsymbol{\alpha})] = \frac{1}{R_l} \operatorname{Tr} \boldsymbol{C}[\rho(\mathbf{s}^{l-1},\boldsymbol{\alpha})].
\end{aligned}$$

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, 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 $\tilde{\mathbf{v}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_1,\mathrm{c}_1},\dots,\tilde{\mathbf{v}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_k,\mathrm{c}_k}$ does not depend on $\boldsymbol{\alpha}$.

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 $\forall \boldsymbol{\alpha}', \mathrm{c} : \quad \tilde{\mathbf{v}}_{\boldsymbol{\alpha}',\mathrm{c}} = \tilde{\mathbf{u}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}',\mathrm{c}}$. 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 mapping Φ_l defined as $(\tilde{\mathbf{x}}^l, \tilde{\mathbf{s}}^l) = \Phi_l(\tilde{\mathbf{x}}, \tilde{\mathbf{s}})$, and thus

$$T_{\boldsymbol{\alpha}}(\tilde{\mathbf{x}}^l, \tilde{\mathbf{s}}^l) = \Phi_l(T_{\boldsymbol{\alpha}}(\tilde{\mathbf{x}}, \tilde{\mathbf{s}})), \tag{32}$$

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 a given 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. (32), 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

$$\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}), \tag{33}$$

$$\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_1,\mathrm{c}_1}, \ldots, \tilde{\mathbf{x}}^l_{\boldsymbol{\alpha}+\boldsymbol{\alpha}_k,\mathrm{c}_k} = \Phi(\tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_1,\mathrm{c}'_1}, \ldots, \tilde{\mathbf{x}}_{\boldsymbol{\alpha}+\boldsymbol{\alpha}'_{k'},\mathrm{c}'_{k'}}), \tag{34}$$

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

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

By stationarity of $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{s}}$, the terms on the right-hand sides of Eq. (33) and (34) have the same distribution, and the terms on the right-hand sides of Eq. (35) and (36) 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 $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \sim_{\mathbf{v},\boldsymbol{\alpha}'} \mathbf{v}_{\boldsymbol{\alpha}',\mathrm{c}}$ for given $\boldsymbol{\alpha}$ and random $\boldsymbol{\alpha}'$, and thus that $RF(\mathbf{v})_{\boldsymbol{\alpha},i_\mathrm{c}} \sim_{\mathbf{v},\boldsymbol{\alpha}} \mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}}$ 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 $\mathbf{s}$ 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 $RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ and $RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:} = RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},:}$ for most $\boldsymbol{\alpha}$. This implies for any index i_c that $P_{\mathbf{x},\boldsymbol{\alpha}}\big[RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big] \simeq P_{\mathbf{x},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big]$ and $P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big] \simeq P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big]$.

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 $P_{\mathbf{x},\boldsymbol{\alpha}}\big[RF^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big] = P_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\big]$

and $P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[RF^{(\mathrm{p})}(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\right] = P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\mathbf{s}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$. We then deduce that $P_{\mathbf{x},\boldsymbol{\alpha}}\left[RF(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\right] \simeq P_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$ and $P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[RF(\mathbf{s}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\right] \simeq P_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\mathbf{s}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$, meaning that RF is approximately statistics-preserving for $\mathbf{x}^{l-1}$ and $\mathbf{s}^{l-1}$.

□

D Details of Section 3

D.1 Assumption that Φ_l is differentiable a.s. with respect to $\mathbf{x}$

Let us detail which parts of Section 3 rely on the assumption that Φ_l is differentiable a.s. with respect to $\mathbf{x}$.

Firstly the *factor of noise equivalence* detailed in Eq. (4) relies on the assumption that moments with respect to $\mathbf{x}, \boldsymbol{\alpha}$ of the true noise tensor $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$ at layer l and $\mathrm{d}\mathbf{x}^l$ defined as the result of the simultaneous propagation of Eq. (1) and Eq. (2) coincide. If $\Phi_l(\mathbf{x})$ is differentiable *a.s.* with respect to $\mathbf{x}$, then $\mathrm{d}\mathbf{x}^l = \Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$ a.s. with respect to $\mathbf{x}$. In that case, $\mathrm{d}\mathbf{x}^l$ and $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$ share the same probability density function, and thus the same moments with respect to $\mathbf{x}, \boldsymbol{\alpha}$

Secondly the *Jabobian equivalence* and the *sensitivity equivalence* detailed respectively in Sections D.3 and D.2 rely on the assumption that $\Phi_l(\mathbf{x})$ is differentiable *surely* with respect to $\mathbf{x}$. This can be relaxed using subdifferentials if $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$. Indeed the same argument as before shows that moments with respect to $\mathbf{x}, \boldsymbol{\alpha}$ are left unchanged when ignoring the probability-zero event such that $\Phi_l(\mathbf{x})$ is not differentiable with respect to $\mathbf{x}$.

Now let us justify the assumption that $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$ in the case of the propagation of Eq. (1). As in Section 4, we adopt the notation $\Theta^l = (\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \ldots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ and we further assume standard initialization. For given $\mathbf{x}$ such that $\forall \boldsymbol{\alpha}$: $\mathbf{x}_{\boldsymbol{\alpha},:} \neq 0$, it is easy to see that $\Phi_l(\mathbf{x}^l)$ is *not differentiable* implies that $\exists k \leq l$, $\exists \boldsymbol{\alpha}, \mathrm{c}$ such that $\rho(\mathbf{x}^{k-1}, \boldsymbol{\alpha}) \neq 0$ and $\mathbf{x}^k_{\boldsymbol{\alpha},\mathrm{c}} = 0$. Under standard initialization, this corresponds to a zero-probability event. Denoting D the event such that $\Phi_l(\mathbf{x}^l)$ is not differentiable, we then have $\mathbb{P}_{\Theta^l|\mathbf{x}}[D] = 0$. Now considering $\mathbf{x}$ again as random, using Fubini's Theorem and making the assumption that $\mathbf{x}_{\boldsymbol{\alpha},:} \neq 0$ a.s. with respect to $\mathbf{x}, \boldsymbol{\alpha}$ (which is the case e.g. if this random vector $\mathbf{x}_{\boldsymbol{\alpha},:}$ admits a well-defined probability density function):

$$\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x}|\Theta^l}\left[\mathbf{1}_D\right] = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\Theta^l|\mathbf{x}}\left[\mathbf{1}_D\right] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{P}_{\Theta^l|\mathbf{x}}[D]\right] = 0. \tag{37}$$

By contradiction, if there would be non-zero probability with respect to Θ^l that $\mathbb{P}_{\mathbf{x}|\Theta^l}[D] > 0$, then Eq. (37) would not hold. Therefore with probability 1 with respect to Θ^l: $\mathbb{P}_{\mathbf{x}|\Theta^l}[D] = 0$, meaning that $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$.

D.2 Property of normalized sensitivity

Proposition 8. *The sensitivity tensor $\mathbf{s}^l$ 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 $\mathrm{d}\mathbf{x}^l$ as a small corruption to $\mathbf{x}^l$, $\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ can be written as a function of $\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and the input noise $\mathrm{d}\mathbf{x}$,

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

where $\langle\cdot,\cdot\rangle$ denotes the standard dot product in input space. It follows from the definition of $\mathbf{s}^l = \mathrm{d}\mathbf{x}^l/\sigma_{\mathrm{d}\mathbf{x}}$ 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

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

□

Proposition 9. *Define $\tilde{\mathbf{x}}$ the rescaling by a constant factor of $\mathbf{x}$ with unit variance $\mu_2(\tilde{\mathbf{x}}) = 1$, and $\tilde{\mathbf{x}}^l$ the rescaling by a constant factor of $\mathbf{x}^l$ with unit variance $\mu_2(\tilde{\mathbf{x}}^l) = 1$. When $\tilde{\mathbf{x}}^l$ is considered*

as a function of $\tilde{\mathbf{x}}$*, the normalized sensitivity* χ^l *measures a root mean square sensitivity as* $\chi^l = \mathbb{E}_{\tilde{\mathbf{x}},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right]^{1/2}$.

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

$$\mu_2(\mathbf{s}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\mathbb{E}_{\mathbf{s}}\left[\hat{\varphi}(\mathbf{s}^l,\boldsymbol{\alpha})_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\mathbb{E}_{\mathbf{s}}\left[\varphi(\mathbf{s}^l,\boldsymbol{\alpha})_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\mathbb{E}_{\mathbf{s}}\left[(\mathbf{s}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] \tag{38}$$

$$= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right], \tag{39}$$

where Eq. (38) follows from $\mathbf{s}^l$ being centered and Eq. (39) follows from Proposition 8. The normalized sensitivity is then given by

$$\chi^l = \left(\frac{\mu_2(\mathbf{s}^l)\,\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)}\right)^{1/2} = \left(\frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right]\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)]}\right)^{1/2}. \tag{40}$$

Let us define $\tilde{\mathbf{x}}^l = \mathbf{x}^l \,/\, \mu_2(\mathbf{x}^l)^{1/2}$ and $\tilde{\mathbf{x}} = \mathbf{x} \,/\, \mu_2(\mathbf{x})^{1/2} = \mathbf{x} \,/\, \mu_2(\mathbf{x}^0)^{1/2}$. Then we get $\nabla_{\mathbf{x}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = \nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} \,/\, \mu_2(\mathbf{x}^l)^{1/2}$ and $\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = \nabla_{\mathbf{x}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}}\,\mu_2(\mathbf{x}^0)^{1/2}$. It follows that

$$\begin{aligned}\mu_2(\tilde{\mathbf{x}}^l) &= 1,\\ \mu_2(\tilde{\mathbf{x}}) &= 1,\\ \mathbb{E}_{\tilde{\mathbf{x}},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right] &= \frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right]\mu_2(\mathbf{x}^0)}{\mu_2(\mathbf{x}^l)]}.\end{aligned} \tag{41}$$

Finally combining Eq. (40) and Eq. (41), we get $\chi^l = \mathbb{E}_{\tilde{\mathbf{x}},\boldsymbol{\alpha},\mathrm{c}}\left[||\nabla_{\tilde{\mathbf{x}}}\tilde{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}}||_2^2\right]^{1/2}$.

□

D.3 Equivalence between χ^l and previous definitions

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

$$\left(\frac{\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right]}{N_l}\frac{\mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^0_{\mathrm{c}}]\right]}{\mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\right]}\right)^{1/2}. \tag{42}$$

Let us prove the equivalence between the definitions of Eq. (3) and Eq. (42). In the fully-connected case $n = 1$, the spatial position $\boldsymbol{\alpha}$ can be ignored and tensors and feature map vectors coincide: $\mathbf{x}^l = \varphi(\mathbf{x}^l)$ and $\mathbf{s}^l = \varphi(\mathbf{s}^l) = \hat{\varphi}(\mathbf{s}^l)$. 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 $\mathrm{d}\mathbf{x}^l = \boldsymbol{J}^l\,\mathrm{d}\mathbf{x}$, and thus the sensitivity $\mathbf{s}^l = \boldsymbol{J}^l\mathbf{s}$. Due to the white noise property $\mathbb{E}_{\mathbf{s}}\left[\mathbf{s}_i\mathbf{s}_j\right] = \delta_{ij}$,

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

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

$$\begin{aligned}\mu_2(\mathbf{s}^l) &= \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\left[\hat{\varphi}(\mathbf{s}^l)_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\left[\varphi(\mathbf{s}^l)_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\mathbf{s},\mathrm{c}}\left[(\mathbf{s}^l_{\mathrm{c}})^2\right],\\ \mu_2(\mathbf{x}^l) &= \mathbb{E}_{\mathbf{x},\mathrm{c}}\left[\hat{\varphi}(\mathbf{x}^l)_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\right],\end{aligned}$$

finally give the equivalence between the two definitions:

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

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

D.4 Distributional pathologies

We consider the following rescaling of the signal:

$$\boldsymbol{\nu}^l \equiv (\nu_{1,\mathrm{c}}(\mathbf{x}^l))_{1\le \mathrm{c}\le N_l}, \qquad \tilde{\mathbf{x}}^l \equiv \frac{1}{||\boldsymbol{\nu}^l||_2}\mathbf{x}^l, \qquad \tilde{\boldsymbol{\nu}}^l \equiv (\nu_{1,\mathrm{c}}(\tilde{\mathbf{x}}^l))_{1\le \mathrm{c}\le N_l}.$$

We immediately have $||\tilde{\boldsymbol{\nu}}^l||_2 = 1$. Furthermore we have

$$\begin{aligned}\nu_2(\mathbf{x}^l) &= \frac{1}{N_l}\sum_{\mathrm{c}} \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}^2\big] = \frac{1}{N_l}\Big(\sum_{\mathrm{c}} \mathrm{Var}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}\big] + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{x}^l,\boldsymbol{\alpha})_{\mathrm{c}}\big]^2\Big)\\ &= \frac{1}{N_l}\Big(\sum_{\mathrm{c}} \mu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{1,\mathrm{c}}(\mathbf{x}^l)^2\Big) = \mu_2(\mathbf{x}^l) + \frac{1}{N_l}||\boldsymbol{\nu}^l||_2^2.\end{aligned}$$

The pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$ then implies $\nu_2(\mathbf{x}^l)/||\boldsymbol{\nu}^l||_2^2 \xrightarrow{l\to\infty} 1/N_l$ and $\mu_2(\mathbf{x}^l)/||\boldsymbol{\nu}^l||_2^2 \xrightarrow{l\to\infty} 0$, i.e. $\mu_2(\tilde{\mathbf{x}}^l) \xrightarrow{l\to\infty} 0$. Thus $\varphi(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})$ becomes *point-like* concentrated around the vector $\tilde{\boldsymbol{\nu}}^l$ of unit L^2 norm.

E Details of Section 5

E.1 Lemma on the sum of increments

Lemma 10. *Consider a sequence (X_k) of random variables which depend on Θ^k and denote $Y_k = \mathbb{E}_{\theta^k}[X_k]$ and $Z_k = X_k - \mathbb{E}_{\theta^k}[X_k]$. Then:*

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

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

(ii) If there exist constants $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$ with $\forall k \le l$: $m_{\min} \le Y_k \le m_{\max}$ and $v_{\min} \le \mathrm{Var}_{\Theta^k}[Z_k] \le v_{\max}$, then there exist random variables m_l and s_l such that s_l is centered and

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

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

$$\begin{aligned}\mathbb{E}_{\theta^k}[Z_k] &= \mathbb{E}_{\theta^k}[X_k] - \mathbb{E}_{\theta^k}[X_k] = 0,\\ \mathbb{E}_{\Theta^k}[Z_k] &= \mathbb{E}_{\Theta^{k-1}}[\mathbb{E}_{\theta^k}[Z_k]] = 0.\end{aligned} \tag{43}$$

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

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

where Eq. (44) follows from Eq. (43).

□

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

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

$$= \sum\nolimits_k \mathbb{E}_{\Theta^k}\left[Z_k^2\right] = \sum\nolimits_k \mathrm{Var}_{\Theta^k}\left[Z_k\right], \tag{45}$$

where Eq. (45) follows from (i). The hypothesis further 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 now define $m_l = M_l \,/\, l$ and $s_l = S_l \,/\, \sqrt{l}$, we get that 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. *There exist positive constants $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max} > 0$, random variables $(m_l), (m'_l)$, $(s_l), (s'_l)$ and events (A_l) with probability $\mathbb{P}_{\Theta^l}[A_l] \geq \prod_{k=1}^l (1 - 2^{-N_k+1})$ such that under A_l: s_l and s'_l are centered and*

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

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

$$\begin{aligned}
\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^l_{\mathrm{c},:}\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right)^2\right] \\
&= \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \lambda_i.
\end{aligned} \tag{46}$$

We further define

$$u^l_{\mathrm{c}} = \begin{cases} \frac{\nu_{2,\mathrm{c}}(\mathbf{x}^l)}{\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)} & \text{if } \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{47}$$

Combining the definition of u^l_{c} with Eq. (24) and Eq. (46), we get that under $\left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$:

$$\begin{aligned}
&\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) = u^l_{\mathrm{c}}\, \nu_{2,\mathrm{c}}(\mathbf{y}^l), \\
&\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) = u^l_{\mathrm{c}} \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \lambda_i = R_l\, \nu_2(\mathbf{x}^{l-1})\, u^l_{\mathrm{c}} \sum\nolimits_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i,
\end{aligned} \tag{48}$$

where we defined $\hat{\lambda}_i = \lambda_i \,/\, \sum_j \lambda_j$ and used $\sum_j \lambda_j = \mathrm{Tr}\, \boldsymbol{G}\left[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right] = R_l \nu_2(\mathbf{x}^{l-1})$ by Corollary 3. The symmetric propagation gives that, under $\left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$:

$$\begin{aligned}
\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= R_l\, \nu_2(\mathbf{x}^{l-1})\, \bar{u}^l_{\mathrm{c}} \sum\nolimits_i \left(-\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i, \\
\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= R_l\, \nu_2(\mathbf{x}^{l-1})\, (1 - u^l_{\mathrm{c}}) \sum\nolimits_i \left(-\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \hat{\lambda}_i,
\end{aligned} \tag{49}$$

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

where $\bar{u}^l_{\mathrm{c}}$ is the symmetric counterpart of u^l_{c} obtained by interverting $\mathbf{x}^l$ and $\bar{\mathbf{x}}^l$ in Eq. (47), and Eq. (49) follows from $\bar{u}^l_{\mathrm{c}} = (1 - u^l_{\mathrm{c}})$ when $\nu_{2,\mathrm{c}}(\mathbf{y}^l) > 0$ and $\bar{u}^l_{\mathrm{c}} \nu_{2,\mathrm{c}}(-\mathbf{y}^l) = (1 - u^l_{\mathrm{c}}) \nu_{2,\mathrm{c}}(-\mathbf{y}^l)$ still holding when $\nu_{2,\mathrm{c}}(\mathbf{y}^l) = 0$. By symmetry of the propagation $\nu_{2,\mathrm{c}}(\mathbf{x}^l) \sim_{\theta^l} \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)$. Combined with Eq. (50) and the assumption of standard initialization $\boldsymbol{W}^l \sim_{\theta^l} \hat{\boldsymbol{W}}^l \sim_{\theta^l} \mathcal{N}(0, \sqrt{2 \,/\, R_l}\boldsymbol{I})$, we deduce that under $\left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$:

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

Thus $\forall \mathrm{c} : \mathbb{E}_{\theta^l}[\nu_{2,\mathrm{c}}(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1})$ and $\mathbb{E}_{\theta^l}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1})$, meaning that under $\left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$:

$$
\mathbb{E}_{\theta^l}[\delta\nu_2(\mathbf{x}^l)] = 1. \tag{51}
$$

Let us define

$$
v^l_{\mathrm{c}} = \begin{cases} 0 & \text{if } (u^l_{\mathrm{c}} < 1/2) \wedge (\nu_{2,\mathrm{c}}(\mathbf{y}^l) > 0) \\ 1 & \text{if } (u^l_{\mathrm{c}} > 1/2) \wedge (\nu_{2,\mathrm{c}}(\mathbf{y}^l) > 0) \\ b & \text{if } (u^l_{\mathrm{c}} = 1/2) \vee (\nu_{2,\mathrm{c}}(\mathbf{y}^l) = 0) \end{cases}
$$

with $\wedge$ the logical *and*, $\vee$ the logical *or*, and with b independent of Θ^l and following a Bernouilli distribution with probability $1/2$: $b \sim \mathrm{Bernouilli}(1/2)$. Conditionally on any value of $\nu_{2,\mathrm{c}}(\mathbf{y}^l)$, we have $v^l_{\mathrm{c}} \sim \mathrm{Bernouilli}(1/2)$ and thus $\nu_{2,\mathrm{c}}(\mathbf{y}^l)$ and v^l_{c} are independent. Let us further denote $B_l = \left\{\exists \mathrm{c} : v^l_{\mathrm{c}} = 1\right\}$, which gives $\mathbb{P}_{\theta^l}[B_l] = 1 - 2^{-N_l}$.

Since $(\nu_{2,\mathrm{c}}(\mathbf{y}^l))_{1\leq \mathrm{c} \leq N_l}$ and $(v^l_{\mathrm{c}})_{1\leq \mathrm{c}\leq N_l}$ are independent, using Eq. (48) we get that $\exists (w_{i,0})$ such that under $B_l \cap \left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$:

$$
(w_{i,0})_{1\leq i\leq R_l} \sim \mathcal{N}(0, \sqrt{2\,/\,R_l}\boldsymbol{I}) : \quad \frac{R_l}{2N_l}\nu_2(\mathbf{x}^{l-1})\sum_{i=1}^{R_l} w^2_{i,0}\hat{\lambda}_i \leq \frac{1}{N_l}\sum_{\mathrm{c}=1}^{R_l}\nu_{2,\mathrm{c}}(\mathbf{x}^l) = \nu_2(\mathbf{x}^l).
$$

Thus $\exists (w_{i,j})_{i,j}$ such that the increments $\delta\nu_2(\mathbf{x}^l)$ can be bounded under $B_l \cap \left\{\nu_2(\mathbf{x}^{l-1}) \neq 0\right\}$ as

$$
\forall j : (w_{i,j})_{1\leq i\leq R_l} \sim \mathcal{N}(0, \sqrt{2\,/\,R_l}\boldsymbol{I}), \quad \frac{R_l}{2N_l}\sum_{i=1}^{R_l} w^2_{i,0}\hat{\lambda}_i \leq \delta\nu_2(\mathbf{x}^l) \leq \frac{R_l}{N_l}\sum_{j=1}^{N_l}\sum_{i=1}^{R_l} w^2_{i,j}\hat{\lambda}_i. \tag{52}
$$

Simply replacing $\mathbf{x}^l$ by $\mathbf{s}^l$, $\mathbf{y}^l$ by $\mathbf{t}^l$, $\boldsymbol{G}$ by $\boldsymbol{C}$, and using Eq. (26) instead of Eq. (24) and the identity with $\mu_2(\mathbf{s}^{l-1})$ instead of $\nu_2(\mathbf{x}^{l-1})$ in Corollary 3, we get that under $\left\{\mu_2(\mathbf{s}^{l-1}) \neq 0\right\}$:

$$
\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{s}^l)] = 1. \tag{53}
$$

We similarly define $B'_l : \mathbb{P}_{\theta^l}\left[B'_l\right] = 1 - 2^{-N_l}$, $(\hat{\lambda}'_i)_{1\leq i\leq R_l} : \sum_i \hat{\lambda}'_i = 1$ and $(w'_{i,j})$ such that, under $B'_l \cap \left\{\mu_2(\mathbf{s}^{l-1}) \neq 0\right\}$:

$$
\forall j : (w'_{i,j})_{1\leq i\leq R_l} \sim \mathcal{N}(0, \sqrt{2\,/\,R_l}\boldsymbol{I}) : \quad \frac{R_l}{2N_l}\sum_{i=1}^{R_l} w'^2_{i,0}\hat{\lambda}'_i \leq \delta\mu_2(\mathbf{s}^l) \leq \frac{R_l}{N_l}\sum_{j=1}^{N_l}\sum_{i=1}^{R_l} w'^2_{i,j}\hat{\lambda}'_i. \tag{54}
$$

Let us denote $A_l = \bigcap_{k=1}^l \left(B_k \cap B'_k \cap \left\{\nu_2(\mathbf{x}^k) \neq 0\right\} \cap \left\{\mu_2(\mathbf{s}^k) \neq 0\right\}\right)$. It follows from Eq. (52) and Eq. (54) that $\left\{\nu_2(\mathbf{x}^k) \neq 0\right\}$ has probability 1 under $B_k \cap A_{k-1}$, and that $\left\{\mu_2(\mathbf{s}^k) \neq 0\right\}$ has probability 1 under $B'_k \cap A_{k-1}$. Thus

$$
\begin{aligned}
\mathbb{P}_{\theta^l|A_{l-1}}\left[A_l\right] &= \mathbb{P}_{\theta^l|A_{l-1}}\left[B_l \cap B'_l \cap \left\{\nu_2(\mathbf{x}^l) \neq 0\right\} \cap \left\{\mu_2(\mathbf{s}^l) \neq 0\right\}\right], \\
&= \mathbb{P}_{\theta^l|A_{l-1}}\left[B_l \cap B'_l\right] = 1 - \mathbb{P}_{\theta^l|A_{l-1}}\left[B^{\mathrm{c}}_l \cup B^{\mathrm{c}\prime}_l\right] \geq 1 - 2^{-N_l+1},
\end{aligned} \tag{55}
$$

where B_k^{c} and $B_k^{\mathrm{c}\prime}$ denote the complementary events of B_k and B_k' respectively. We then have for $\mathbb{P}_{\Theta^l}\left[A_l\right]$:

$$\begin{aligned}\mathbb{P}_{\Theta^l}\left[A_l\right] &= \prod_{k=1}^{l} \mathbb{P}_{\theta^k|A_{k-1}}\left[A_k\right] \\ &\geq \prod_{k=1}^{l}\left(1-2^{-N_k+1}\right).\end{aligned}$$

From Eq. (55), we deduce that $\mathbb{P}_{\theta^l|A_{l-1}}\left[A_l^{\mathrm{c}}\right] \ll 1$, where A_l^{c} denotes the complementary event of A_l. Further using $\mathbb{E}_{\theta^l|A_{l-1}}\left[\delta\nu_2(\mathbf{x}^l)\right] = 1$, and $\mathbb{E}_{\theta^l|A_{l-1}}\left[\delta\mu_2(\mathbf{s}^l)\right] = 1$ by Eq. (51) and Eq. (53), and applying Cauchy-Schwarz inequality:

$$\begin{aligned}&\left|\mathbb{E}_{\theta^l|A_l}\left[\delta\nu_2(\mathbf{x}^l)\right]-1\right| \\ &= \left|\mathbb{E}_{\theta^l|A_l}\left[\delta\nu_2(\mathbf{x}^l)\right]-\mathbb{E}_{\theta^l|A_{l-1}}\left[\delta\nu_2(\mathbf{x}^l)\right]\right| \\ &= \left|\left(\mathbb{P}_{\theta^l|A_{l-1}}[A_l]^{-1}-1\right)\mathbb{E}_{\theta^l|A_{l-1}}\left[\mathbf{1}_{A_l}\delta\nu_2(\mathbf{x}^l)\right]+\mathbb{E}_{\theta^l|A_{l-1}}\left[\mathbf{1}_{A_l^{\mathrm{c}}}\delta\nu_2(\mathbf{x}^l)\right]\right| \\ &\leq \mathbb{P}_{\theta^l|A_{l-1}}[A_l^{\mathrm{c}}]\mathbb{P}_{\theta^l|A_{l-1}}[A_l]^{-1}\left|\mathbb{E}_{\theta^l|A_{l-1}}\left[\mathbf{1}_{A_l}\delta\nu_2(\mathbf{x}^l)\right]\right|+\mathbb{P}_{\theta^l|A_{l-1}}\left[A_l^{\mathrm{c}}\right]^{1/2}\mathbb{E}_{\theta^l|A_{l-1}}\left[\delta\nu_2(\mathbf{x}^l)^2\right]^{1/2} \\ &\leq \left(\frac{\mathbb{P}_{\theta^l|A_{l-1}}[A_l^{\mathrm{c}}]}{\mathbb{P}_{\theta^l|A_{l-1}}[A_l]}+\mathbb{P}_{\theta^l|A_{l-1}}\left[A_l^{\mathrm{c}}\right]^{1/2}\right)\mathbb{E}_{\theta^l|A_{l-1}}\left[\delta\nu_2(\mathbf{x}^l)^2\right]^{1/2} \ll 1. \qquad (56)\end{aligned}$$

Similarly,

$$\left|\mathbb{E}_{\theta^l|A_l}\left[\delta\mu_2(\mathbf{s}^l)\right]-1\right| \ll 1. \qquad (57)$$

By log-concavity, it follows that $\mathbb{E}_{\theta^l|A_l}\left[\log\delta\nu_2(\mathbf{x}^l)\right] \leq 0$, and that $\mathbb{E}_{\theta^l|A_l}\left[\log\delta\mu_2(\mathbf{s}^l)\right] \leq 0$. Now since $\log x$ and $\log^2 x$ are integrable at 0, it follows from Eq. (52) and Eq. (54) that $\delta\nu_2(\mathbf{x}^l)$ has well-defined first and second-order moments with respect to θ^l under $A_{l-1}\cap B_l\cap\left\{\nu_2(\mathbf{x}^l)\neq 0\right\}$, and that $\delta\mu_2(\mathbf{s}^l)$ has well-defined first and second-order moments with respect to θ^l under $A_{l-1}\cap B_l'\cap\left\{\mu_2(\mathbf{s}^l)\neq 0\right\}$. Thus $\delta\nu_2(\mathbf{x}^l)$ and $\delta\mu_2(\mathbf{s}^l)$ have well-defined first and second-order moments with respect to θ^l under A_l.

From now on we work under A_l, and we define under A_l:

$$\begin{aligned}X_k &= \log\delta\nu_2(\mathbf{x}^k), \quad Y_k = \mathbb{E}_{\theta^k|A_l}[\log\delta\nu_2(\mathbf{x}^k)], \quad Z_k = \log\delta\nu_2(\mathbf{x}^k)-\mathbb{E}_{\theta^k|A_l}[\log\delta\nu_2(\mathbf{x}^k)], \\ X_k' &= \log\delta\mu_2(\mathbf{s}^k), \quad Y_k' = \mathbb{E}_{\theta^k|A_l}[\log\delta\mu_2(\mathbf{s}^k)], \quad Z_k' = \log\delta\mu_2(\mathbf{s}^k)-\mathbb{E}_{\theta^k|A_l}[\log\delta\mu_2(\mathbf{s}^k)].\end{aligned}$$

Note that for $k<l$, A_l is independent of Θ^k under A_k. This implies

$$\begin{aligned}Y_k &= \mathbb{E}_{\theta^k|A_l}[\log\delta\nu_2(\mathbf{x}^k)] = \mathbb{E}_{\theta^k|A_k}[\log\delta\nu_2(\mathbf{x}^k)] \leq 0, \qquad \mathrm{Var}_{\Theta^k|A_k}[Z_k] = \mathrm{Var}_{\Theta^k|A_l}[Z_k], \\ Y_k' &= \mathbb{E}_{\theta^k|A_l}[\log\delta\mu_2(\mathbf{s}^k)] = \mathbb{E}_{\theta^k|A_k}[\log\delta\mu_2(\mathbf{s}^k)] \leq 0, \qquad \mathrm{Var}_{\Theta^k|A_k}[Z_k'] = \mathrm{Var}_{\Theta^k|A_l}[Z_k].\end{aligned}$$

Let us now apply Lemma 10 for given l conditionally on A_l. Suppose there exist positive constants $m_{\min}, m_{\max}, v_{\min}, v_{\max} > 0$, such that $\forall k \leq l$ under A_l:

$$m_{\min} \leq -Y_k, -Y_k' \leq m_{\max}, \qquad v_{\min} \leq \mathrm{Var}_{\theta^k|A_k}[Z_k], \mathrm{Var}_{\theta^k|A_k}[Z_k'] \leq v_{\max},$$

implying

$$-m_{\max} \leq Y_k, Y_k' \leq -m_{\min}, \qquad v_{\min} \leq \mathrm{Var}_{\Theta^k|A_l}[Z_k], \mathrm{Var}_{\Theta^k|A_l}[Z_k'] \leq v_{\max}.$$

Then by Lemma 10 there exist sequences of random variables $(m_l), (m_l'), (s_l), (s_l')$ such that $\forall l$ under A_l: s_l and s_l' are centered and

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

where we used $\mu_2(\mathbf{s}^0) = 1$ due to the white noise property: $\mathbb{E}_{\mathbf{s}}[\mathbf{s}_i \mathbf{s}_j] = \delta_{ij}$. By changing the variable m_l to $-m_l$ and the variable m'_l to $-m'_l$, we get that $\forall l$ under A_l: s_l and s'_l are centered and

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

To obtain the bounds $m_{\text{min}}, m_{\text{max}}, v_{\text{min}}, v_{\text{max}}$, we consider extreme cases for u_{c}^l and $R_l \sum_i (\hat{\boldsymbol{W}}_{\text{c},i}^l)^2 \hat{\lambda}_i$ in Eq. (48). Let us denote Chi-Square(N) the chi-square distribution with N degree of freedom and $N_{\text{min}}, N_{\text{max}}, R_{\text{min}}, R_{\text{max}}$ the bounds on N_l, R_l such that

$$N_{\text{min}} = \min_l N_l, \quad N_{\text{max}} = \max_l N_l, \quad R_{\text{min}} = \min_l R_l, \quad R_{\text{max}} = \max_l R_l.$$

Then we obtain *minimum bounds* by considering $N_l = N_{\text{max}}$, $R_l = R_{\text{max}}$ and $u_{\text{c}}^l \sim_{\theta^l} 1/2$ and $R_l \sum_i (\hat{\boldsymbol{W}}_{\text{c},i}^l)^2 \hat{\lambda}_i \sim_{\theta^l} 2\,\text{Chi-Square}(R_{\text{max}}) \,/\, R_{\text{max}}$. This leads to $\log \delta\nu_2(\mathbf{x}^l), \log \delta\mu_2(\mathbf{s}^l) \sim_{\theta^l} \text{Chi-Square}(N_{\text{max}} R_{\text{max}}) \,/\, (N_{\text{max}} R_{\text{max}})$. We obtain *maximum bounds* by considering $N_l = N_{\text{min}}$, and $u_{\text{c}}^l \sim_{\theta^l} \text{Bernouilli}(1/2)$ and $R_l \sum_i (\hat{\boldsymbol{W}}_{\text{c},i}^l)^2 \hat{\lambda}_i \sim_{\theta^l} 2\,\text{Chi-Square}(1)$.

As an illustration, in the fully-connected case with constant width $N_l = 100$, we numerically find $m_{\text{min}} \simeq 9.7{\times}10^{-5}$ and $v_{\text{min}} \simeq 2.0{\times}10^{-4}$ as minimum bounds and $m_{\text{max}} \simeq 2.5{\times}10^{-2}$ and $v_{\text{max}} \simeq 5.2{\times}10^{-2}$ as maximum bounds.

□

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

Here we relate Theorem 1 to the terms $\overline{m}$, $\underline{m}$, $\underline{s}$ defined in Section 4, under the conditionality A_k. By Eq. (56) and Eq. (57), we have $|\mathbb{E}_{\theta^k | A_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \ll 1$ and $|\mathbb{E}_{\theta^k | A_k}[\delta\mu_2(\mathbf{s}^k)] - 1| \ll 1$, which implies

$$|\overline{m}[\nu_2(\mathbf{x}^k)]| = \left|\log \mathbb{E}_{\theta^k | A_k}[\delta\nu_2(\mathbf{x}^k)]\right| \simeq \left|\mathbb{E}_{\theta^k | A_k}[\delta\nu_2(\mathbf{x}^k)] - 1\right| \ll 1,$$
$$|\overline{m}[\mu_2(\mathbf{s}^k)]| \ll 1.$$

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 $\underline{m}[\nu_2(\mathbf{x}^k)] < 0$, $\underline{m}[\mu_2(\mathbf{s}^k)] < 0$. These terms correspond to Y_k, Y'_k in the proof of Theorem 1.

E.4 Proof of Theorem 2

Theorem 2. Normalized Sensitivity increments of vanilla networks. *Denote* $\mathbf{y}^{l,+} = \max(\mathbf{y}^l, 0)$ *and* $\mathbf{y}^{l,-} = \max(-\mathbf{y}^l, 0)$. *The dominating term under* A_l *in the evolution of* χ^l *is*

$$\delta\chi^l \simeq \exp\left(\overline{m}_{\text{vanilla}}[\chi^l]\right) = \left(1 - \mathbb{E}_{\text{c},\theta^l | A_{l-1}}\left[\frac{\nu_{1,\text{c}}(\mathbf{y}^{l,+})\nu_{1,\text{c}}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-1/2}. \tag{58}$$

Proof. The dominating term in the evolution of χ^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: $\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{s}^l)] = 1$, 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 $\boldsymbol{C}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^l = \boldsymbol{W}^l(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$. Using these notations, we get

$$\begin{aligned}\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\hat{\varphi}(\mathbf{y}^l, \boldsymbol{\alpha})_{\mathrm{c}}^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^l_{\mathrm{c},:}\hat{\rho}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\right)^2\right] \\ &= \sum_i \left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2 \lambda_i. \end{aligned} \tag{59}$$

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

$$\begin{aligned}\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)] &= \frac{2}{R_l}\sum_i \lambda_i = \frac{2}{R_l}\operatorname{Tr}\boldsymbol{C}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})] \\ &= 2\mu_2(\mathbf{x}^{l-1}). \end{aligned} \tag{60}$$

where Eq. (60) follows from Corollary 3. Furthermore the symmetric propagation gives:

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2\right] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}\right]^2 + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2\right] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}\right]^2. \\ &= \nu_{2,c}(\mathbf{y}^{l,+}) - \nu_{1,c}(\mathbf{y}^{l,+})^2 + \nu_{2,c}(\mathbf{y}^{l,-}) - \nu_{1,c}(\mathbf{y}^{l,-})^2 \\ &= \nu_{2,c}(\mathbf{y}^l) - \left(\nu_{1,c}(\mathbf{y}^{l,+})^2 + \nu_{1,c}(\mathbf{y}^{l,-})^2\right) \end{aligned} \tag{61}$$

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. (61) as

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= \nu_{2,c}(\mathbf{y}^l) - \nu_{1,c}(\mathbf{y}^l)^2 - 2\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-}) \\ &= \mu_{2,c}(\mathbf{y}^l) - 2\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-}) \end{aligned} \tag{62}$$

Combining Eq. (60) and Eq. (62):

$$\begin{aligned}\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l|A_{l-1}}[\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})], \\ 2\mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l|A_{l-1}}[\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})], \\ \mathbb{E}_{\theta^l|A_{l-1}}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= \mu_2(\mathbf{x}^{l-1})\left(1 - \mathbb{E}_{\theta^l|A_{l-1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{l,+})\nu_{1,c}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right). \end{aligned} \tag{63}$$

where Eq. (63) is obtained by 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 $\mathbb{E}_{\theta^l|A_{l-1}}[\delta\mu_2(\mathbf{s}^l)] = 1$:

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

□

E.5 If χ^l has exponential drift larger than diffusion and $\mu_2(\mathbf{x}^l)$, $\nu_2(\mathbf{x}^l)$ are lognormal, then $\mu_2(\mathbf{x}^l)\,/\,\nu_2(\mathbf{x}^l)$ converges a.s. to zero

Lemma 11. Borel Cantelli implies a.s. convergence *(see Durrett (1996))* **.** *For a sequence of random variables (X_l) and a random variable X, if $\forall \epsilon > 0 : \sum_{l=1}^{\infty} \mathbb{P}[|X_l - X| > \epsilon] < \infty$, then*

$$X_l \xrightarrow{l\to\infty} X \text{ a.s.}$$

Proof. For given $\epsilon > 0$, denote N_ϵ the number of times that the event $\{|X_l - X| > \epsilon\}$ occurs such that $N_\epsilon = \sum_{l=1}^{\infty} \mathbf{1}_{\{|X_l - X| > \epsilon\}}$. Fubini's Theorem implies $\mathbb{E}[N_\epsilon] = \sum_{l=1}^{\infty} \mathbb{P}[|X_l - X| > \epsilon] < \infty$, and thus that N_ϵ is finite a.s.

Now let us reason by contradiction and suppose $\exists E$ with $\mathbb{P}[E] > 0$ such that, under E: $X_l \xrightarrow{l\to\infty\!\!\!\!\!\!/} X$. Under E, $\exists \epsilon$ random variable and $\exists (k_l)$ increasing sequence with $\forall l$: $|X_{k_l} - X| > \epsilon$. This implies in turn $\exists \epsilon' > 0$ non-random and $\exists E'$ with $\mathbb{P}[E'] > 0$ such that under E': $\exists (k_l)$ increasing sequence with $\forall l$: $|X_{k_l} - X| > \epsilon'$. Thus $N_{\epsilon'}$ thus has non-zero probability to be infinite: $\mathbb{P}[N_{\epsilon'} = \infty] \geq \mathbb{P}[E'] > 0$, which is a contradiction. We deduce that $X_l \xrightarrow{l\to\infty} X$ a.s.

□

Proposition 12. *Suppose that*

(i) We can neglect the events A_l with probability exponentially small in the width

(ii) The event D under which χ^l has drift larger than diffusion has probability $\mathbb{P}[D] > 0$

(iii) $\mu_2(\mathbf{x}^l)$, $\nu_2(\mathbf{x}^l)$ are lognormal

Then, under D:

$$\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} \xrightarrow{l\to\infty} 0 \text{ a.s.} \tag{64}$$

Proof. Neglecting the events A_l, Theorem 1 implies that $\exists (m_l), (m'_l), (s_l), (s'_l)$ such that $(s_l), (s'_l)$ are centered and

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

Given that $\log \nu_2(\mathbf{x}^l)$ and $\log \mu_2(\mathbf{s}^l)$ are Gaussian by the assumption of lognormality, and given $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^0)$ and $\mathbb{E}_{\Theta^l}[\mu_2(\mathbf{s}^l)] = \mu_2(\mathbf{s}^0) = 1$ under standard initialization, we deduce: $\exists (X_l), (X'_l)$ random variables and $\exists (M_l), (M'_l) > 0$ constants such that

$$\begin{aligned}
&\log \nu_2(\mathbf{x}^l) = X_l - M_l + \log \nu_2(\mathbf{x}^0), \quad && X_l \sim \mathcal{N}(0, \sqrt{M_l}), \quad && l m_{\min} \leq M_l \leq l m_{\max},\\
&\log \mu_2(\mathbf{s}^l) = X'_t - M'_l, && X'_l \sim \mathcal{N}(0, \sqrt{M'_l}), \quad && l m_{\min} \leq M'_l \leq l m_{\max}.
\end{aligned}$$

Now let us make more precise the conditionality on D. We may assume $\exists m > (m_{\max} - m_{\min})/2$ such that $\forall l$ under D: $\log \chi^l \geq l m$.

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

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

which gives with logarithms that, under D:

$$\begin{aligned}
\log \mu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^l) &= -2 \log \chi^l + \log \mu_2(\mathbf{s}^l) - \log \nu_2(\mathbf{x}^l) + \log \mu_2(\mathbf{x}^0)\\
&\leq -2 l m + \big(X'_l - M'_l\big) - \big(X_l - M_l + \log \nu_2(\mathbf{x}^0)\big) + \log \mu_2(\mathbf{x}^0)
\end{aligned}$$

$$\leq -2lm + lm_{\max} - lm_{\min} - \log \nu_2(\mathbf{x}^0) + \log \mu_2(\mathbf{x}^0) + X'_l - X_l$$
$$\leq -lM + C + X'_l - X_l,$$

with $M = 2m - m_{\max} + m_{\min} > 0$ and $C = -\log \nu_2(\mathbf{x}^0) + \log \mu_2(\mathbf{x}^0)$. Thus for given ϵ, under D:

$$\begin{aligned}
\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon \implies & \log \epsilon < -lM + C + X'_l - X_l \\
\implies & \Big(X'_l \geq \frac{\log \epsilon + lM - C}{2}\Big) \vee \Big(-X_l \geq \frac{\log \epsilon + lM - C}{2}\Big) \\
\implies & \Big(\tilde{X}'_l \geq \frac{\log \epsilon + lM - C}{2\sqrt{M'_l}}\Big) \vee \Big(-\tilde{X}_l \geq \frac{\log \epsilon + lM - C}{2\sqrt{M_l}}\Big) \\
\implies & \Big(\tilde{X}'_l \geq \frac{\log \epsilon + lM - C}{2\sqrt{lm_{\max}}}\Big) \vee \Big(-\tilde{X}_l \geq \frac{\log \epsilon + lM - C}{2\sqrt{lm_{\max}}}\Big),
\end{aligned}$$

with $\vee$ the logical *or*, and with $\tilde{X}_l \equiv X_l/\sqrt{M_l}$ and $\tilde{X}'_l \equiv X'_l/\sqrt{M'_l}$. In turn, this implies $\exists C_\epsilon$ such that $\forall l$, under D:

$$\begin{aligned}
\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon \implies & \Big(\tilde{X}'_l \geq \sqrt{l}C_\epsilon\Big) \vee \Big(-\tilde{X}_l \geq \sqrt{l}C_\epsilon\Big), \\
\mathbb{P}_D\left[\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon\right] &\leq \mathbb{P}_D\big[\tilde{X}'_l \geq \sqrt{l}C_\epsilon\big] + \mathbb{P}_D\big[-\tilde{X}_l \geq \sqrt{l}C_\epsilon\big] \\
&\leq \frac{1}{\mathbb{P}[D]}\mathbb{P}\Big[1_D \mathbf{1}_{\{\tilde{X}'_l \geq \sqrt{l}C_\epsilon\}}\Big] + \frac{1}{\mathbb{P}[D]}\mathbb{P}\Big[1_D \mathbf{1}_{\{-\tilde{X}_l \geq \sqrt{l}C_\epsilon\}}\Big] \\
&\leq \frac{1}{\mathbb{P}[D]}\operatorname{erfc}\left(\sqrt{\frac{l}{2}}C_\epsilon\right) \qquad (65) \\
&\leq \frac{1}{\mathbb{P}[D]}\exp\Big(-\frac{l}{2}C_\epsilon^2\Big), \qquad (66)
\end{aligned}$$

where Eq. (65) is obtained using $\tilde{X}_l, \tilde{X}'_l \sim \mathcal{N}(0,1)$, while Eq. (66) is obtained using $\operatorname{erfc}(x) \leq \exp(-x^2)$ (Chiani et al., 2003). It follows from Eq. (66) that

$$\sum_{l=1}^{\infty} \mathbb{P}_D\left[\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon\right] < \infty.$$

By Lemma 11, we finally deduce that, under D:

$$\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} \xrightarrow{l\to\infty} 0 \text{ a.s.}$$

□

E.6 If $\exp\left(\overline{m}[\chi^l]\right) \to 1$ and $\tilde{\mathbf{x}}^l$ has bounded moments, then $\mathbf{x}^l$ converges to one-dimensional signal pathology

Proposition 13. *We recall the notation for the unit-variance rescaled signal: $\tilde{\mathbf{x}}^l \equiv \mathbf{x}^l/\mu_2(\mathbf{x}^l)^{1/2}$, and the usual notation:*

$$X_l = \mathcal{O}(Y_l) \iff \exists M > 0, \forall l : |X_l| \leq M|Y_l|.$$

We further suppose that

- *(i) $\tilde{\mathbf{x}}^l$ is well-defined with bounded moments: $\nu_p(|\tilde{\mathbf{x}}^l|) = \mathcal{O}(1)$, implying in particular $\nu_2(\mathbf{x}^l)/\mu_2(\mathbf{x}^l) \not\xrightarrow{l\to\infty} \infty$ and thus $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \not\xrightarrow{l\to\infty} 0$, i.e. that $\mathbf{x}^l$ does not converge to zero-dimensional signal pathology,*

(ii) $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \xrightarrow{l\to\infty} 1$.

Then $\mathbf{x}^l$ *converges to one-dimensional signal pathology.*

Proof. Again we use the notations from Section B and we denote

$$\boldsymbol{\nu}_\varphi^l \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})\big] = (\nu_{1,\mathrm{c}}(\tilde{\mathbf{x}}^l))_{1\le \mathrm{c}\le N_l}, \qquad \boldsymbol{\nu}_\rho^l \equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})\big].$$

Due to the statistic-preserving property, we have $\frac{1}{N_l}||\boldsymbol{\nu}_\varphi^l||_2^2 = \frac{1}{R_l}||\boldsymbol{\nu}_\rho^l||_2^2$, which implies

$$\begin{aligned}\nu_2(\tilde{\mathbf{x}}^l) &= \frac{1}{N_l}\Big(\sum_{\mathrm{c}}\mu_{2,\mathrm{c}}(\tilde{\mathbf{x}}^l) + \nu_{1,\mathrm{c}}(\tilde{\mathbf{x}}^l)^2\Big) = \mu_2(\tilde{\mathbf{x}}^l) + \frac{1}{N_l}||\boldsymbol{\nu}_\varphi^l||_2^2 \\ &= 1 + \frac{1}{N_l}||\boldsymbol{\nu}_\varphi^l||_2^2 = 1 + \frac{1}{R_l}||\boldsymbol{\nu}_\rho^l||_2^2\end{aligned}$$

i.e. $||\boldsymbol{\nu}_\rho^l||_2^2 = R_l\big(\nu_2(\tilde{\mathbf{x}}^l) - 1\big)$. Combined with $\nu_2(\tilde{\mathbf{x}}^l) = \mathcal{O}(1)$, we deduce that $||\boldsymbol{\nu}_\rho^l||_2 = \mathcal{O}(1)$.

Now let us reason by contradiction and suppose that $r_{\text{eff}}(\mathbf{x}^l) = r_{\text{eff}}(\tilde{\mathbf{x}}^l) \xrightarrow{l\to\infty\;\not}\; 1$, which implies $\exists\eta > 0$ and $\exists(k_l)$ increasing sequence with $\forall l$: $r_{\text{eff}}(\tilde{\mathbf{x}}^{k_l}) \ge 1+\eta$. In turn this implies $\exists\eta' > 0$ such that $\forall l$:

$$\exists \mathbf{v}_\varphi^{k_l} \in \mathbb{R}^{N_{k_l}} \perp \boldsymbol{\nu}_\varphi^{k_l}, \quad ||\mathbf{v}_\varphi^{k_l}||_2 = 1: \quad \mathrm{Var}_{\mathbf{x},\boldsymbol{\alpha}}\big[\langle\varphi(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\varphi^{k_l}\rangle\big] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\langle\varphi(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\varphi^{k_l}\rangle^2\big] \ge \eta',$$

i.e. that $\varphi(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})$ necessarily has a direction of variance $> \eta'$ which is orthogonal to its mean vector $\boldsymbol{\nu}_\varphi^{k_l}$. By padding this direction appropriately with zeros, it follows that $\exists\eta' > 0$ such that $\forall l$:

$$\exists \mathbf{v}_\rho^{k_l} \in \mathbb{R}^{R_{k_l}} \perp \boldsymbol{\nu}_\rho^{k_l}, \quad ||\mathbf{v}_\rho^{k_l}||_2 = 1: \quad \mathrm{Var}_{\mathbf{x},\boldsymbol{\alpha}}\big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\rho^{k_l}\rangle\big] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\rho^{k_l}\rangle^2\big] \ge \eta'.$$

Let us denote $\tilde{\boldsymbol{W}}^{k_l+1}$ such that $\forall \mathrm{c}: \tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1} = \boldsymbol{W}_{\mathrm{c},:}^{k_l+1}/||\boldsymbol{W}_{\mathrm{c},:}^{k_l+1}||_2$ and $\tilde{\boldsymbol{\nu}}_\rho^{k_l} = \boldsymbol{\nu}_\rho^{k_l}/||\boldsymbol{\nu}_\rho^{k_l}||_2$. We further decompose $\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}$ as

$$\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1} = w_{\mathbf{v}}\mathbf{v}_\rho^{k_l} + w_{\boldsymbol{\nu}}\tilde{\boldsymbol{\nu}}_\rho^{k_l} + \mathbf{w}', \qquad \mathbf{w}' \perp \mathbf{v}_\rho^{k_l}, \qquad \mathbf{w}' \perp \tilde{\boldsymbol{\nu}}_\rho^{k_l}.$$

Then we get

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2 &\ge \Big(w_{\mathbf{v}}^2\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\rho^{k_l}\rangle^2\big]\Big)^2 \ge w_{\mathbf{v}}^4\eta'^2, \\ \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2 &= w_{\boldsymbol{\nu}}^2||\boldsymbol{\nu}_\rho^{k_l}||_2^2.\end{aligned}$$

Given that $||\boldsymbol{\nu}_\rho^{k_l}||_2 = \mathcal{O}(1)$, this implies by spherical symmetry that $\forall\epsilon > 0$, $\exists p_\epsilon > 0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\left[\Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2 \ge \eta'^2 - \epsilon\Big) \wedge \Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2 \le \epsilon\Big)\right] \ge p_\epsilon, \tag{67}$$

with $\wedge$ the logical *and*. By Cauchy-Schwarz inequality:

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2 \le \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big]\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|^3\Big]. \tag{68}$$

We can bound the second term on the right-hand side as

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}_{\mathrm{c},:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|^3\Big]$$

$$\leq \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\Big(\sum_{i_c=1}^{R_l}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_c}^2\Big)^{3/2}\right] \tag{69}$$

$$\leq \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\sum_{i_1,i_2,i_3}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_1}^2\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_2}^2\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_3}^2\right]^{1/2} \tag{70}$$

$$\leq \sum_{i_1,i_2,i_3}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_1}^4\right]^{1/4}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_2}^8\right]^{1/8}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_3}^8\right]^{1/8} \tag{71}$$

$$\leq R_l^3 N_l^{1/2}\nu_4(\tilde{\mathbf{x}}^{k_l})^{1/4}\nu_8(\tilde{\mathbf{x}}^{k_l})^{1/4}, \tag{72}$$

where Eq. (69), Eq. (70) and Eq. (71) are obtained by again applying Cauchy-Schwarz inequality, while Eq. (72) is obtained with $\forall i, \forall p$: $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^p\right] \leq \sum_c \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\varphi(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_c^p\right] = N_l\nu_p(\tilde{\mathbf{x}}^l)$. It then follows from Eq. (68) and the hypothesis that all moments are bounded $\nu_p(|\tilde{\mathbf{x}}^l|) = \mathcal{O}(1)$ that

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\tilde{\boldsymbol{W}}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^2\right]^2 = \mathcal{O}\Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big]\Big). \tag{73}$$

Combining Eq. (67) and Eq. (73), we deduce that $\exists \eta'' > 0$ with $\forall \epsilon > 0$, $\exists p_\epsilon > 0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\left[\left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\big|\tilde{\boldsymbol{W}}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\right] \geq \eta'' - \epsilon\right) \wedge \left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\tilde{\boldsymbol{W}}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right]^2 \leq \epsilon\right)\right] \geq p_\epsilon.$$

Due to the assumption of standard initialization $\boldsymbol{W}^{k_l+1} \sim_{\theta^{k_l+1}} \mathcal{N}(0, \sqrt{2\,/\,R_{k_l}}\boldsymbol{I})$, it follows that $\tilde{\boldsymbol{W}}^{k_l+1}$ and $||\boldsymbol{W}_{c,:}^{k_l+1}||_2$ are independent, and we have e.g. $\mathbb{P}_{\theta^{k_l+1}}\left[1 \leq ||\boldsymbol{W}_{c,:}^{k_l+1}||_2 \leq 2\right] \geq 0$ independent of l. Therefore $\forall \epsilon > 0$, $\exists p'_\epsilon > 0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\left[\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\big|\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\right] \geq \eta'' - \epsilon\right) \wedge \left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right]^2 \leq 4\epsilon\right)\right] \geq p'_\epsilon. \tag{74}$$

Let us note that

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\big|\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^+\right] + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^-\right],$$

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right]^2 = \left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^+\right] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^-\right]\right)^2,$$

where $x^+ \equiv \max(x, 0)$ and $x^- \equiv \max(-x, 0)$. We then get $\exists \eta''' > 0$, $\exists p > 0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\left[\left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^+\right] \geq \eta'''\right) \wedge \left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}_{c,:}^{k_l+1}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\right)^-\right] \geq \eta'''\right)\right] \geq p,$$

$$\mathbb{P}_{\theta^{k_l+1}}\left[\left(\nu_{1,c}(\mathbf{y}^{k_l+1,+}) \geq \eta'''\mu_2(\mathbf{x}^l)^{1/2}\right) \wedge \left(\nu_{1,c}(\mathbf{y}^{k_l+1,-}) \geq \eta'''\mu_2(\mathbf{x}^l)^{1/2}\right)\right] \geq p,$$

$$\mathbb{P}_{\theta^{k_l+1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{k_l+1,+})\nu_{1,c}(\mathbf{y}^{k_l+1,-})}{\mu_2(\mathbf{x}^l)} \geq (\eta''')^2\right] \geq p.$$

We finally get

$$\mathbb{E}_{c,\theta^{k_l+1}}\left[\frac{\nu_{1,c}(\mathbf{y}^{k_l+1,+})\nu_{1,c}(\mathbf{y}^{k_l+1,-})}{\mu_2(\mathbf{x}^l)}\right] \geq p(\eta''')^2.$$

Thus by Theorem 2, $\exists \eta'''' > 0$ such that $\forall l$:

$$\exp\left(\overline{m}[\chi^{k_l+1}]\right) \geq 1 + \eta'''',$$

which contradicts the hypothesis $\exp\left(\overline{m}[\chi^{k_l+1}]\right) \xrightarrow{l\to\infty} 1$. We deduce that $r_{\text{eff}}(\mathbf{x}^l) \xrightarrow{l\to\infty} 1$, i.e. that $\mathbf{x}^l$ converges to one-dimensional signal pathology.

□

E.7 If $\exp\left(\overline{m}[\chi^l]\right) \to 1$, then each new layer l becomes arbitrary well approximated by a linear function

If $\exp\left(\overline{m}[\chi^l]\right) \to 1$, then Eq. (9) in Theorem 2 implies for the rescaled signal $\tilde{\mathbf{y}}^l = \mathbf{y}^l / \sqrt{\mu_2(\mathbf{x}^{l-1})}$:

$$\begin{aligned}
\mathbb{E}_{\mathrm{c},\theta^l}\left[\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right] &\to 0,\\
\mathbb{E}_{\mathrm{c},\theta^l}\left[\min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right)^2\right] &\to 0,\\
\forall \epsilon: \quad \mathbb{P}_{\mathrm{c},\theta^l}\left[\min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) \geq \epsilon\right] &\to 0,\\
\forall \epsilon: \quad \mathbb{P}_{\theta^l}\left[\exists \mathrm{c}: \ \min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) \geq \epsilon\right] &\to 0,\\
\forall \epsilon: \quad \mathbb{P}_{\theta^l}\left[\forall \mathrm{c}: \ \min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) \leq \epsilon\right] &\to 1.
\end{aligned} \tag{75}$$

Now let us fix a channel c and suppose that $\min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) \leq \epsilon$. Given that $\tilde{\mathbf{y}}^{l,-} = \tilde{\mathbf{y}}^l - \tilde{\mathbf{y}}^{l,+}$, we have

$$\min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) = \min\left(\nu_{1,\mathrm{c}}(|\tilde{\mathbf{y}}^{l,+} - 0|), \nu_{1,\mathrm{c}}(|\tilde{\mathbf{y}}^{l,+} - \tilde{\mathbf{y}}^l|)\right) \leq \epsilon.$$

Both $\nu_{1,\mathrm{c}}(|\tilde{\mathbf{y}}^{l,+} - 0|)$ and $\nu_{1,\mathrm{c}}(|\tilde{\mathbf{y}}^{l,+} - \tilde{\mathbf{y}}^l|)$ correspond to the mean absolute error of the approximation of the rescaled output $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1}) = \mathbf{y}^{l,+}/\mu_2(\mathbf{x}^{l-1}) = \tilde{\mathbf{y}}^{l,+}$ in channel c with a linear function of $\mathbf{x}^{l-1}$. So there exists a linear function $f_{\mathrm{c}} : \mathbb{R}^{n\times\cdots\times n\times N_{l-1}} \to \mathbb{R}^{n\times\cdots\times n}$ such that

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[|\tilde{\mathbf{y}}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}} - f_{\mathrm{c}}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha}}|\right] \leq \epsilon.$$

If $\forall \mathrm{c}$: $\min\left(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\right) \leq \epsilon$, and if we define the linear function $f : \mathbb{R}^{n\times\cdots\times n\times N_{l-1}} \to \mathbb{R}^{n\times\cdots\times n\times N_l}$ such that $\forall \boldsymbol{\alpha}, \mathrm{c}$: $f(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},\mathrm{c}} = f_{\mathrm{c}}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha}}$, then we get

$$\nu_1(|\tilde{\mathbf{y}}^{l,+} - f(\mathbf{x}^{l-1})|) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\left[|\tilde{\mathbf{y}}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}} - f(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},\mathrm{c}}|\right] = \mathbb{E}_{\mathrm{c}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[|\tilde{\mathbf{y}}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}} - f_{\mathrm{c}}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha}}|\right] \leq \epsilon.$$

Combined with Eq. (75), this means that $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1})$ can be approximated arbitrary well by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrary close to 1 in θ^l.

F Details of Section 6

F.1 Proof of Theorem 3

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

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

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

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

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

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

Term $\exp\left(\overline{m}_{BN}\left[\chi^l\right]\right)$**.** First 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 $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)]$:

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

All together, we get

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

Term $\exp\left(\overline{m}_{\phi}\left[\chi^l\right]\right)$**.** We consider the symmetric propagation for batch-normalized feedforward nets. From Eq. (31) we deduce

$$\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)] + \mathbb{E}_{\theta^l}[\mu_2(\bar{\mathbf{s}}^l)] = \mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)],$$
$$2\mathbb{E}_{\theta^l}[\mu_2(\mathbf{s}^l)] = \mathbb{E}_{\theta^l}[\mu_2(\mathbf{u}^l)], \tag{76}$$

where Eq. (76) is obtained by symmetry of the propagation. Now we turn to the symmetric propagation of the signal:

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2\right] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}\right]^2 + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[(\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2\right] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}\right]^2. \tag{77}$$
$$= \nu_{2,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{2,c}(\mathbf{z}^{l,-}) - \nu_{1,c}(\mathbf{z}^{l,-})^2$$
$$= \nu_{2,c}(\mathbf{z}^l) - \left(\nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2\right),$$

where Eq. (77) follows from Eq. (29). 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

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = 1 - \left(\nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2\right). \tag{78}$$
$$\nu_{1,c}(\mathbf{z}^l) = \nu_{1,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,-}) = 0,$$
$$\left(\nu_{1,c}(\mathbf{z}^{l,+}) - \nu_{1,c}(\mathbf{z}^{l,-})\right)^2 = \nu_{1,c}(\mathbf{z}^{l,+})^2 + \nu_{1,c}(\mathbf{z}^{l,-})^2 - 2\nu_{1,c}(\mathbf{z}^{l,+})\nu_{1,c}(\mathbf{z}^{l,-}) = 0. \tag{79}$$

Using Eq. (78), Eq. (79) and the symmetry of the propagation,

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = 1 - 2\nu_{1,c}(\mathbf{z}^{l,+})\nu_{1,c}(\mathbf{z}^{l,-}),$$
$$2\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)] = 1 - 2\mathbb{E}_{\mathrm{c},\theta^l}\left[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})\right]. \tag{80}$$

We finally combine Eq. (76) and Eq. (80):

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

□

F.2 In the first step of the propagation: $\exp\left(\overline{m}_{BN}\left[\chi^1\right]\right) \geq 1$

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

$$\mu_{2,\mathrm{c}}(\mathbf{t}^1) = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\hat{\varphi}(\mathbf{t}^1,\boldsymbol{\alpha})_\mathrm{c}^2\right] = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\varphi(\mathbf{t}^1,\boldsymbol{\alpha})_\mathrm{c}^2\right] = \mathbb{E}_{\mathbf{x},\mathbf{s},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^1_{\mathrm{c},:}\rho(\mathbf{s},\boldsymbol{\alpha})\right)^2\right] \tag{81}$$
$$= \sum\nolimits_{i,j} \boldsymbol{W}^1_{\mathrm{c},i}\boldsymbol{W}^1_{\mathrm{c},j}\mathbb{E}_{\mathbf{s},\boldsymbol{\alpha}}[\rho(\mathbf{s},\boldsymbol{\alpha})_i\rho(\mathbf{s},\boldsymbol{\alpha})_j] = \sum\nolimits_i \left(\boldsymbol{W}^1_{\mathrm{c},i}\right)^2 = ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2. \tag{82}$$

where Eq. (81) follows from $\mathbf{t}^1$ being centered and Eq. (82) 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 $\mathbb{E}_\mathbf{s}[\rho(\mathbf{s},\boldsymbol{\alpha})_i\rho(\mathbf{s},\boldsymbol{\alpha})_j] = \delta_{ij}$ under periodic boundary conditions.

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

$$\mu_{2,\mathrm{c}}(\mathbf{y}^1) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\hat{\varphi}(\mathbf{y}^1,\boldsymbol{\alpha})_\mathrm{c}^2\right] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^1_{\mathrm{c},:}\hat{\rho}(\mathbf{x},\boldsymbol{\alpha})\right)^2\right] = \sum\nolimits_i \left(\hat{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i$$
$$= ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i = \mu_{2,\mathrm{c}}(\mathbf{t}^1)\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i,$$

where we defined $\tilde{\boldsymbol{W}}^1$ such that $\forall \mathrm{c}$: $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:} = \hat{\boldsymbol{W}}^1_{\mathrm{c},:} \,/\, ||\boldsymbol{W}^1_{\mathrm{c},:}||$ and we used Eq. (82). 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

$$\forall i:\ \mathbb{E}_{\theta^1}\left[\left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\right] = \frac{1}{R_1},$$
$$\forall \mathrm{c}:\ \mathbb{E}_{\theta^1}\left[\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i\right] = \frac{1}{R_1}\sum\nolimits_i \lambda_i, \qquad \mathbb{E}_{\mathrm{c},\theta^1}\left[\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i\right] = \frac{1}{R_1}\sum\nolimits_i \lambda_i. \tag{83}$$

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

$$\exp\left(\overline{m}_{BN}\left[\chi^1\right]\right) = \left(\frac{\mu_2(\mathbf{s}^0)}{\mu_2(\mathbf{x}^0)}\right)^{-1/2}\mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^1)}{\mu_{2,\mathrm{c}}(\mathbf{y}^1)}\right]^{1/2}$$
$$= \left(\frac{1}{R_1}\sum_i \lambda_i\right)^{1/2}\mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{1}{\sum_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i}\right]^{1/2}, \tag{84}$$
$$\geq \left(\frac{1}{R_1}\sum_i \lambda_i\right)^{1/2}\left(\mathbb{E}_{\mathrm{c},\theta^1}\left[\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i\right]^{-1}\right)^{1/2} = 1. \tag{85}$$

where Eq. (84) is obtained using $\mu_2(\mathbf{s}^0) = \mu_2(\mathbf{s}) = 1$ and $\mu_2(\mathbf{x}^0) = \frac{1}{R_1}\operatorname{Tr}\boldsymbol{C}[\rho(\mathbf{x},\boldsymbol{\alpha})] = \frac{1}{R_1}\sum_i \lambda_i$ by Corollary 3, while Eq. (85) is obtained using the convexity of $x \mapsto 1/x$ and Eq. (83).

G Details of Section 7

G.1 Adaptation of the previous setup to resnets

Before proceeding to the analysis, slight adaptations and forewords are necessary. Let us denote $\Theta^{l,h} = (\boldsymbol{\omega}^{1,1},\boldsymbol{\beta}^{1,1},\dots,\boldsymbol{\omega}^{1,H},\boldsymbol{\beta}^{1,H},\dots,\boldsymbol{\omega}^{l,1},\boldsymbol{\beta}^{l,1},\dots,\boldsymbol{\omega}^{l,h},\boldsymbol{\beta}^{l,h})$ the full set of parameters up to

layer h in residual unit l and $\theta^{l,h} = \Theta^{l,h}|\Theta^{l,h-1}$ 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.

In the pre-activation perspective of Eq. (16) and (17) each residual layer starts with $(\mathbf{y}^{l,h-1}, \mathbf{t}^{l,h-1})$ after the convolution and ends with $(\mathbf{y}^{l,h}, \mathbf{t}^{l,h})$ again after the convolution. The concrete effect is that batch normalization and ϕ are completely deterministic conditionally on Θ^{l-1} in the first layer $h = 1$ of each residual unit l. 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 $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ of each residual unit becomes more and more correlated between successive l and less and less dependent on the parameters θ^{l-k} of previous individual units.

Since the evolution of χ^l is mainly influenced by batch normalization and the nonlinearity ϕ, this shift can be thought as attributing the parameters and thus the stochasticity of layer h to layer $h-1$. A simple strategy to apply the results of Section 6 is thus to shift back to the post-activation perspective by considering the parameters $\theta^{l,h-1}$ and the evolution from $(\mathbf{x}^{l,h-1}, \mathbf{s}^{l,h-1})$ to $(\mathbf{x}^{l,h}, \mathbf{s}^{l,h})$ for layers $2 \leq h \leq H$. Theorem 3 strictly applies in this case.

It remains to understand the evolution from $(\mathbf{y}^{l,0}, \mathbf{t}^{l,0}) = (\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ to $(\mathbf{x}^{l,1}, \mathbf{s}^{l,1})$ in layer $h = 1$ and the evolution from $(\mathbf{x}^{l,H}, \mathbf{s}^{l,H})$ to $(\mathbf{y}^{l,H}, \mathbf{t}^{l,H})$ in layer $h = H$. By considering the parameter Θ^{l-1} and the evolution from $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1})$ to $(\mathbf{x}^{l,1}, \mathbf{s}^{l,1})$, the batch normalization term of Eq. (12) in Theorem 3 becomes

$$\left(\frac{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{u}^{l,1})]}{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{t}^{l-1})]}\right)^{1/2} \left(\frac{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{z}^{l,1})]}{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{y}^{l-1})]}\right)^{-1/2}$$
$$= \left(\frac{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{t}^{l-1})]}{\mathbb{E}_{\Theta^{l-1}}[\mu_2(\mathbf{y}^{l-1})]}\right)^{-1/2} \mathbb{E}_{\mathrm{c},\Theta^{l-1}}\left[\frac{\mu_{2,\mathrm{c}}(\mathbf{t}^{l-1})}{\mu_{2,\mathrm{c}}(\mathbf{y}^{l-1})}\right]^{1/2}.$$

Under the assumption of well-conditioned sensitivity $\mu_{2,\mathrm{c}}(\mathbf{t}^l) \simeq \mu_{2,\mathrm{c}}(\mathbf{t}^{l-1})$, this term is again $\gtrsim 1$ by convexity of $x \mapsto 1/x$. For the nonlinearity term, symmetric propagation with respect to Θ^{l-1} applies for all terms in the sum $(\mathbf{y}^{l-1}, \mathbf{t}^{l-1}) = \sum_{k=0}^{l-1} (\mathbf{y}^{k,H}, \mathbf{t}^{k,H})$ except for $(\mathbf{y}^{0,H}, \mathbf{t}^{0,H}) = (\mathbf{y}, \mathbf{t})$. The expression of the nonlinearity term of Eq. (13) in Theorem 3 thus remains approximately valid.

Finally by spherical symmetry, the evolution from $(\mathbf{x}^{l,H}, \mathbf{s}^{l,H})$ to $(\mathbf{y}^{l,H}, \mathbf{t}^{l,H})$ in layer $h = H$ has dominating term

$$\left(\frac{\mathbb{E}_{\theta^{l,H}}[\mu_2(\mathbf{t}^{l,H})]}{\mathbb{E}_{\theta^{l,H}}[\mu_2(\mathbf{s}^{l,H})]}\right)^{1/2} \left(\frac{\mathbb{E}_{\theta^{l,H}}[\mu_2(\mathbf{y}^{l,H})]}{\mathbb{E}_{\theta^{l,H}}[\mu_2(\mathbf{x}^{l,H})]}\right)^{-1/2} = 1.$$

In summary, Theorem 3 remains approximately valid in the feedforward evolution inside residual units.

G.2 Lemma on dot-product

Lemma 14. *The following hold:*

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0,$$
$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]^2\right] \leq \frac{1}{N r_{\mathrm{eff}}(\mathbf{y}^{l-1})}\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{y}^{l,H})],$$
$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\mathbf{t},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{t}^{l-1})_{\mathrm{c}}\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0,$$
$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\mathbf{t},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{t}^{l-1})_{\mathrm{c}}\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]^2\right] \leq \frac{1}{N r_{\mathrm{eff}}(\mathbf{y}^{l-1})}\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}[\mu_2(\mathbf{t}^{l,H})].$$

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

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = \mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\big(-\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\big)\right]\right],$$
$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]\right] = 0.$$

Next we note that

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right] &= \frac{1}{N}\sum_{\mathrm{c}}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right],\\ &= \frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right], \qquad (86)\end{aligned}$$

where $\langle\cdot\rangle$ denotes the standard dot product in $\mathbb{R}^N$. Let us denote $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$ and $(\lambda_1,\dots,\lambda_N)$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})]$. We further denote u_i the unit-variance components of $\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})$ in the basis $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$, and y_i the components of $\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})$ in the basis $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$. This gives

$$\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha}) = \sum_i \sqrt{\lambda_i}u_i\boldsymbol{e}_i, \quad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[u_i\right] = 0, \quad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[u_i^2\right] = 1,$$
$$\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha}) = \sum_i y_i\boldsymbol{e}_i.$$

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

$$\forall j:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[z_iu_j\right] = 0, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_iu_i\right] = \alpha_{i,i}, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_i^2\right] = \sum_j \alpha_{i,j}^2 + \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[z_i^2\right],$$
$$\mu_2(\mathbf{y}^{l,H}) = \frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\langle\mathbf{y}^{l,H},\mathbf{y}^{l,H}\rangle\right] = \frac{1}{N}\sum_i \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_i^2\right] = \frac{1}{N}\Big(\sum_{i,j}\alpha_{i,j}^2 + \sum_i \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[z_i^2\right]\Big). \qquad (87)$$

where the dot product in Eq. (87) is computed in the orthogonal basis $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$. Now computing the dot product of $\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})$ and $\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})$ in the orthogonal basis $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$:

$$\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right] = \sum_i \sqrt{\lambda_i}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}[y_iu_i] = \sum_i \sqrt{\lambda_i}\alpha_{i,i}.$$

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. Thus:

$$\begin{aligned}&\forall j\neq i:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_iu_i\right]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_ju_j\right] \sim_{\theta^l} \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[-y_iu_i\right]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[y_ju_j\right],\\ &\forall j\neq i:\ \alpha_{i,i}\alpha_{j,j} \sim_{\theta^l} (-\alpha_{i,i})\alpha_{j,j},\\ &\forall j\neq i:\ \mathbb{E}_{\theta^l}\left[\alpha_{i,i}\alpha_{j,j}\right] = 0.\end{aligned}$$

We deduce that

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right]^2\right] = \sum_i \lambda_i\mathbb{E}_{\theta^l}\left[\alpha_{i,i}^2\right].$$

Spherical symmetry also implies that the distribution of $\alpha_{i,j}$ with respect to θ^l does not depend on i. Denoting (β_j) such that $\forall i,j$: $\beta_j = \mathbb{E}_{\theta^l}\left[\alpha_{i,j}^2\right]$, we get combined with Eq. (87):

$$\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right] \geq \frac{1}{N}\sum_{i,j}\mathbb{E}_{\theta^l}\left[\alpha_{i,j}^2\right] = \frac{1}{N}\sum_{i,j}\beta_j = \sum_i \beta_i,$$

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right]^2\right] = \sum_i \lambda_i\beta_i \leq \lambda_{\max}\Big(\sum_i \beta_i\Big) \leq \lambda_{\max}\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right].$$

Finally combining with Eq. (86):

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

where we used $\lambda_{\max}r_{\text{eff}}(\mathbf{y}^{l-1}) = \sum_i \lambda_i = N\mu_2(\mathbf{y}^{l-1})$. The same analysis can be applied to $\hat{\varphi}(\mathbf{t}^{l-1},\boldsymbol{\alpha})$ and $\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})$.

□

Corollary 15. *Let us denote the dot products:*

$$\begin{aligned}Y_l &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right],\\ T_l &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{t}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right],\\ Y_{l,l} &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right] = \mu_2(\mathbf{y}^{l,H}),\\ T_{l,l} &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right] = \mu_2(\mathbf{t}^{l,H}).\end{aligned}$$

Then by spherical symmetry:

$$\begin{aligned}&\forall l:\ \mathbb{E}_{\Theta^l}[Y_l] = 0, \qquad \forall l \neq l':\ \mathbb{E}_{\Theta^{\max(l,l')}}[Y_l Y_{l'}] = 0,\\ &\forall l:\ \mathbb{E}_{\Theta^l}[T_l] = 0, \qquad \forall l \neq l':\ \mathbb{E}_{\Theta^{\max(l,l')}}[T_l T_{l'}] = 0.\end{aligned}$$

Furthermore given $r_{\text{eff}}(\mathbf{y}^{l-1})$, $r_{\text{eff}}(\mathbf{t}^{l-1}) \geq 1$ and given Lemma 14, we deduce the following inequalities in preparation of the proof of Theorem 4:

$$\begin{aligned}\mathbb{E}_{\Theta^l}\left[Y_l^2\right] &\leq \frac{1}{N}\mathbb{E}_{\Theta^{l-1}}\Big[\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right]\Big],\\ \mathbb{E}_{\Theta^l}\left[\left(\frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}T_l\right)^2\right] &\leq \frac{1}{N}\mathbb{E}_{\Theta^{l-1}}\left[\frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^{l-1})}{(\chi^{l-1})^2}\mathbb{E}_{\theta^l}\left[\frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^{l,H})}{(\chi^{l-1})^2}\right]\right],\\ &\leq \frac{1}{N}\mathbb{E}_{\Theta^{l-1}}\left[\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^{l,H})}{(\chi^{l-1})^2}\right]\right].\end{aligned}$$

G.3 Proof of Theorem 4

Theorem 4. Normalized Sensitivity increments of batch-normalized resnets. *Suppose that for all depth l we can bound the second-order central moments $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and the feedforward increments inside residual units $\delta_{\min} \lesssim \delta\chi^{l,h} \lesssim \delta_{\max}$. Denote $\eta_{\min} = \left((\delta_{\min})^{2H}\mu_{2,\min} - \mu_{2,\max}\right)/\mu_{2,\max}$ and $\eta_{\max} = \left((\delta_{\max})^{2H}\mu_{2,\max} - \mu_{2,\min}\right)/\mu_{2,\min}$, as well as $\tau_{\min} = \eta_{\min}/2$ and $\tau_{\max} = \eta_{\max}/2$. Then there exist positive constants $C_{\min}$, $C_{\max} > 0$ such that*

$$\left(1 + \frac{\eta_{\min}}{l+1}\right)^{1/2} \lesssim \delta\chi^l \lesssim \left(1 + \frac{\eta_{\max}}{l+1}\right)^{1/2}, \tag{88}$$

$$C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}. \tag{89}$$

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

We also remind that we write $a \lesssim b$ when $a(1+\epsilon_a) \leq b(1+\epsilon_b)$ with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$ with high probability. And we write $a \simeq b$ when $a(1+\epsilon_a) = b(1+\epsilon_b)$ with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$ with high probability. Denoting $\wedge$ the logical *and*, $\vee$ the logical *or*, the following rules are easily verified:

$$\begin{aligned}
(a \lesssim b) \wedge (a \gtrsim b) &\iff (a \simeq b),\\
(a \lesssim b) &\iff (-a \gtrsim -b),\\
(a \lesssim b) \wedge (c \lesssim d) &\implies (ac \lesssim bd),\\
(a \lesssim b) \wedge (b \lesssim c) &\implies (a \lesssim c),\\
(a \lesssim b) \wedge (a > 0) \wedge (b > 0) &\implies (\sqrt{a} \lesssim \sqrt{b}),\\
(a \lesssim b) \wedge (a > 0) \wedge (b > 0) &\implies (1/a \gtrsim 1/b),\\
(a \lesssim b) \wedge (c \lesssim d) \wedge \Big(\mathbb{P}\Big[\big(|a+c| \ll |a| + |c|\big) \vee \big(|b+d| \ll |b| + |d|\big)\Big] \ll 1\Big)&\\
\implies (a + c \lesssim b + d).&
\end{aligned}$$

Finally consider a constant b and a random variable a depending on Θ^l with well-defined moments. Let us prove that $(a \lesssim b) \implies \big(\mathbb{E}_{\theta^l}[a] \lesssim b\big) \wedge \big(\mathbb{E}_{\Theta^l}[a] \lesssim b\big)$. Denoting A the event under which $a(1+\epsilon_a) \leq b(1+\epsilon_b)$ with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$, we can write

$$\frac{1}{\mathbb{E}_{\theta^l}[a]^2}\Big(\mathbb{E}_{\theta^l}[a] - \mathbb{E}_{\theta^l}[\mathbf{1}_A a]\Big)^2 = \frac{1}{\mathbb{E}_{\theta^l}[a]^2}\mathbb{E}_{\theta^l}[\mathbf{1}_{A^{\mathrm{c}}} a]^2 \leq \mathbb{P}_{\theta^l}[\mathbf{1}_{A^{\mathrm{c}}}]\frac{\mathbb{E}_{\theta^l}[a^2]}{\mathbb{E}_{\theta^l}[a]^2}, \tag{90}$$

where A^{c} denotes the complementary event of A and Eq. (90) is obtained using Cauchy-Schwarz inequality. The assumption $(a \lesssim b)$ further gives $\mathbb{P}_{\Theta^l}[A] \simeq 1$ and $\mathbb{P}_{\Theta^l}[A^{\mathrm{c}}] \ll 1$. By contradiction if we had non negligible probability with respect to Θ^{l-1} that $\mathbb{P}_{\theta^l}[\mathbf{1}_{A^{\mathrm{c}}}] = \mathbb{P}_{\Theta^l|\Theta^{l-1}}[\mathbf{1}_{A^{\mathrm{c}}}]$ is non negligible, then we would not have $\mathbb{P}_{\Theta^l}[A^{\mathrm{c}}] = \mathbb{E}_{\Theta^{l-1}}\mathbb{E}_{\Theta^l|\Theta^{l-1}}[\mathbf{1}_{A^{\mathrm{c}}}] \ll 1$. It follows that $\mathbb{P}_{\theta^l}[\mathbf{1}_{A^{\mathrm{c}}}] \ll 1$ with high probability with respect to Θ^{l-1} and that $\mathbb{E}_{\theta^l}[a] \simeq \mathbb{E}_{\theta^l}[\mathbf{1}_A a] \lesssim b$.

A similar reasoning gives

$$\frac{1}{\mathbb{E}_{\Theta^l}[a]^2}\Big(\mathbb{E}_{\Theta^l}[a] - \mathbb{E}_{\Theta^l}[\mathbf{1}_A a]\Big)^2 \leq \mathbb{P}_{\Theta^l}[\mathbf{1}_{A^{\mathrm{c}}}]\frac{\mathbb{E}_{\Theta^l}[a^2]}{\mathbb{E}_{\Theta^l}[a]^2}, \qquad \mathbb{E}_{\Theta^l}[a] \simeq \mathbb{E}_{\Theta^l}[\mathbf{1}_A a] \lesssim b.$$

We keep all these rules in mind in the course of this proof.

Proof of Eq. (88). Adopting the same notations as Corollary 15 and using $\mathbf{y}^l = \mathbf{y}^{l-1} + \mathbf{y}^{l,H}$ by Eq. (15), we get

$$\begin{aligned}
\mu_2(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\Big(\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}} + \hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\Big)^2\right] = \mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2Y_l, \qquad (91)\\
\mu_2(\mathbf{t}^l) &= \mathbb{E}_{\mathbf{y},\mathbf{t},\boldsymbol{\alpha},\mathrm{c}}\left[\Big(\hat{\varphi}(\mathbf{t}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}} + \hat{\varphi}(\mathbf{t}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\Big)^2\right] = \mu_2(\mathbf{t}^{l-1}) + T_{l,l} + 2T_l.
\end{aligned}$$

Due to $\mu_{2,\min} \lesssim Y_{l,l} = \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$, we have $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^0) = \mu_2(\mathbf{y}^{0,H}) \lesssim \mu_{2,\max}$.

Now let us reason by induction and suppose that $l\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l-1}) \lesssim l\mu_{2,\max}$. Combined with Eq. (91) and Corollary 15 we deduce that

$$\begin{aligned}
l\mu_{2,\min} + \mu_{2,\min} + 2Y_l \lesssim \mu_2(\mathbf{y}^l) \lesssim l\mu_{2,\max} + \mu_{2,\max} + 2Y_l,&\\
\mathbb{E}_{\Theta^l}\left[Y_l^2\right] \lesssim \frac{1}{N} l\mu_{2,\max}^2 \leq \frac{1}{N}\frac{1}{l+1}(l+1)^2\mu_{2,\max}^2.&
\end{aligned}$$

Further using Chebyshev's inequality, we deduce that

$$\mathbb{P}_{\Theta^l}\left[\left|Y_l\right| > k\frac{1}{\sqrt{N}}\frac{1}{\sqrt{l+1}}(l+1)\mu_{2,\max}\right] \lesssim \frac{1}{k^2}. \tag{92}$$

For large width $N \gg 1$, it follows that $\left|Y_l\right| \ll (l+1)\mu_{2,\min}$ and $\left|Y_l\right| \ll (l+1)\mu_{2,\max}$ with high probability, which then gives

$$(l+1)\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^l) \lesssim (l+1)\mu_{2,\max}. \tag{93}$$

Eq. (93) then holds for all l. Now let us write $(\chi^l)^2$ as

$$(\chi^l)^2 = \frac{\mu_2(\mathbf{y}^0)\mu_2(\mathbf{t}^l)}{\mu_2(\mathbf{y}^l)} = \mu_2(\mathbf{y}^0)\frac{\mu_2(\mathbf{t}^{l-1}) + T_{l,l} + 2T_l}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2Y_l},$$
$$(\chi^l)^2 = (\chi^{l-1})^2\,\frac{\mu_2(\mathbf{y}^{l-1}) + \frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}T_{l,l} + 2\frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}T_l}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2Y_l}.$$

Denoting $\tilde{T}_{l,l} = \frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}T_{l,l}$ and $\tilde{T}_l = \frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}T_l$, we then get

$$(\delta\chi^l)^2 = \frac{(\chi^l)^2}{(\chi^{l-1})^2} = \frac{\mu_2(\mathbf{y}^{l-1}) + \tilde{T}_{l,l} + 2\tilde{T}_l}{\mu_2(\mathbf{y}^{l-1}) + Y_{l,l} + 2Y_l}. \tag{94}$$

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

$$\mu_{2,\min} \lesssim Y_{l,l} \lesssim \mu_{2,\max}, \tag{95}$$
$$\tilde{T}_{l,l} = \frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}\mu_2(\mathbf{t}^{l,H}) = \frac{\mu_2(\mathbf{y}^0)}{(\chi^{l-1})^2}(\chi^{l-1})^2\prod\nolimits_h(\delta\chi^{l,h})^2\frac{\mu_2(\mathbf{y}^{l,H})}{\mu_2(\mathbf{y}^0)},$$
$$\gamma_{\min}\mu_{2,\min} \lesssim \tilde{T}_{l,l} \lesssim \gamma_{\max}\mu_{2,\max}. \tag{96}$$

By Corollary 15, variances of Y_l and $\tilde{T}_l$ are bounded as

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

The reasoning of Eq. (92) then implies that $\left|Y_l\right| \ll \mu_2(\mathbf{y}^{l-1})$ and $\left|\tilde{T}_l\right| \ll \mu_2(\mathbf{y}^{l-1})$ with high probability. Finally combining Eq. (94), Eq. (95) and Eq. (96):

$$\frac{\mu_2(\mathbf{y}^{l-1}) + \gamma_{\min}\mu_{2,\min}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\max}} \lesssim (\delta\chi^l)^2 \lesssim \frac{\mu_2(\mathbf{y}^{l-1}) + \gamma_{\max}\mu_{2,\max}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\min}},$$
$$1 + \frac{\gamma_{\min}\mu_{2,\min} - \mu_{2,\max}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\max}} \lesssim (\delta\chi^l)^2 \lesssim 1 + \frac{\gamma_{\max}\mu_{2,\max} - \mu_{2,\min}}{\mu_2(\mathbf{y}^{l-1}) + \mu_{2,\min}},$$
$$1 + \frac{\gamma_{\min}\mu_{2,\min} - \mu_{2,\max}}{(l+1)\mu_{2,\max}} \lesssim (\delta\chi^l)^2 \lesssim 1 + \frac{\gamma_{\max}\mu_{2,\max} - \mu_{2,\min}}{(l+1)\mu_{2,\min}},$$
$$\left(1 + \frac{\eta_{\min}}{l+1}\right)^{1/2} \lesssim \delta\chi^l \lesssim \left(1 + \frac{\eta_{\max}}{l+1}\right)^{1/2}.$$

□

Proof of Eq. (89). Expanding Eq. (88), we get

$$\prod_{k=1}^{l}\left(1 + \frac{\eta_{\min}}{k+1}\right)^{1/2} \lesssim \chi^l = \prod_{k=1}^{l}\delta\chi^k \lesssim \prod_{k=1}^{l}\left(1 + \frac{\eta_{\max}}{k+1}\right)^{1/2}.$$

We can further explicitate the bounds:

$$\begin{aligned}
&\sum_{k=1}^{l} \log\left(1+\frac{\eta_{\max}}{k+1}\right) \\
&\quad \leq \int_1^{l+1} \log\left(1+\frac{\eta_{\max}}{x}\right)dx, \\
&\quad \leq \int_1^{l+1} \log(x+\eta_{\max})dx - \int_1^{l+1} \log x dx, \\
&\quad \leq \Big[x\log x - x\Big]_{1+\eta_{\max}}^{l+1+\eta_{\max}} - \Big[x\log x - x\Big]_1^{l+1}, \\
&\quad \leq (l+1+\eta_{\max})\log(l+1+\eta_{\max}) - (1+\eta_{\max})\log(1+\eta_{\max}) - (l+1)\log(l+1), \\
&\quad \leq \eta_{\max}\log(l+1+\eta_{\max}) + (l+1)\log\left(1+\frac{\eta_{\max}}{l+1}\right) - (1+\eta_{\max})\log(1+\eta_{\max}), \\
&\quad \leq \eta_{\max}\log(l+1+\eta_{\max}) + \eta_{\max} - (1+\eta_{\max})\log(1+\eta_{\max}), \qquad (97)
\end{aligned}$$

where we used $\log(1+x) \leq x$ in Eq. (97). Considering the integration between 2 and $l+2$, we get with a similar calculation:

$$\begin{aligned}
&\sum_{k=1}^{l} \log\left(1+\frac{\eta_{\min}}{k+1}\right) \\
&\quad \geq \eta_{\min}\log(l+2+\eta_{\min}) + (l+2)\log\left(1+\frac{\eta_{\min}}{l+2}\right) - (2+\eta_{\min})\log(2+\eta_{\min}) + 2\log 2, \\
&\quad \geq \eta_{\min}\log(l+2+\eta_{\min}) - (2+\eta_{\min})\log(2+\eta_{\min}) + 2\log 2.
\end{aligned}$$

Let $c_{\max} = \exp\Big(\eta_{\max} - (1+\eta_{\max})\log(1+\eta_{\max})\Big)$, $c_{\min} = \exp\Big(-(2+\eta_{\min})\log(2+\eta_{\min}) + 2\log 2\Big)$. We get

$$\begin{aligned}
\prod_{k=1}^{l}\left(1+\frac{\eta_{\max}}{k+1}\right) &\leq c_{\max}(l+1+\eta_{\max})^{\eta_{\max}}, \\
\prod_{k=1}^{l}\left(1+\frac{\eta_{\min}}{k+1}\right) &\geq c_{\min}(l+2+\eta_{\min})^{\eta_{\min}},
\end{aligned}$$

$$\begin{aligned}
\sqrt{c_{\min}}(l+2+\eta_{\min})^{\eta_{\min}/2} \lesssim \chi^l \lesssim \sqrt{c_{\max}}(l+1+\eta_{\max})^{\eta_{\max}/2}, \\
\sqrt{c_{\min}}(l+2+\eta_{\min})^{\tau_{\min}} \lesssim \chi^l \lesssim \sqrt{c_{\max}}(l+1+\eta_{\max})^{\tau_{\max}}.
\end{aligned}$$

Since $x \mapsto \left(\frac{x+2+\eta_{\min}}{x}\right)^{\tau_{\min}}$ and $x \mapsto \left(\frac{x+1+\eta_{\max}}{x}\right)^{\tau_{\max}}$ are lower-bounded and upper-bounded for $x \geq 1$, there exist positive constants $C_{\min}, C_{\max} > 0$ such that

$$C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}.$$

□