# Characterizing Well-Behaved vs. Pathological Deep Neural Networks

**Antoine Labatie** [1]

## Abstract

We introduce a novel 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 easily incorporates batch normalization and skip-connections. Our key insight is to consider the evolution with depth of statistical moments of signal and noise, thereby characterizing the presence or absence of pathologies in the hypothesis space encoded by the choice of hyperparameters. We establish: (i) for feedforward networks, with and without batch normalization, the multiplicativity of layer composition inevitably leads to ill-behaved moments and pathologies; (ii) for residual networks with batch normalization, on the other hand, skip-connections induce power-law rather than exponential behaviour, leading to well-behaved moments and no pathology.

## 1. Introduction

The feverish pace of practical applications has led in the recent years to many advances in neural network architectures, initialization and regularization. At the same time, theoretical research has not been able to follow the same pace. In particular, there is still no mature theory able to validate the full choices of hyperparameters leading to state-of-the-art performance. This is unfortunate since such theory could also serve as a guide towards further improvement.

Amidst the research aimed at building this theory, an important branch has focused on networks at initialization. Due to the randomness of model parameters at initialization, characterizing networks at that time can be seen as characterizing the hypothesis space of input-output mappings that will be favored or reachable during training, i.e. the inductive bias encoded by the choice of hyperparameters. This view has received strong experimental support, with well-behaved input-output mappings at initialization extensively found to be predictive of trainability and post-training performance (Schoenholz et al., 2017; Yang & Schoenholz, 2017; Xiao et al., 2018; Philipp & Carbonell, 2018; Yang et al., 2019).

Yet, even this simplifying case of networks at initialization is challenging as it notably involves dealing with: (i) the complex interplay of the randomness from input data and from model parameters; (ii) the broad spectrum of potential pathologies; (iii) the finite number of units in each layer; (iv) the difficulty to incorporate convolutional layers, batch normalization and skip-connections. Complexities (i), (ii) typically lead to restricting to specific cases of input data and pathologies, e.g. exploding complexity of data manifolds (Poole et al., 2016; Raghu et al., 2017), exponential correlation or decorrelation of two data points (Schoenholz et al., 2017; Balduzzi et al., 2017; Xiao et al., 2018), exploding and vanishing gradients (Yang & Schoenholz, 2017; Philipp et al., 2018; Hanin, 2018; Yang et al., 2019), exploding and vanishing activations (Hanin & Rolnick, 2018). Complexity (iii) commonly leads to making simplifying assumptions, e.g. convergence to Gaussian processes for infinite width (Neal, 1996; Roux & Bengio, 2007; Lee et al., 2018; Matthews et al., 2018; Borovykh, 2018; Garriga-Alonso et al., 2019; Novak et al., 2019; Yang, 2019), "typical" activation patterns (Balduzzi et al., 2017). Finally complexity (iv) most often leads to limiting the number of hard-to-model elements incorporated at a time. To the best of our knowledge, all attempts have thus far been limited in either their scope or their simplifying assumptions.

As the first contribution of this paper, we introduce a novel approach for the characterization of deep neural networks at initialization. This approach: (i) offers a unifying treatment of the broad spectrum of pathologies without any restriction on the input data; (ii) requires only mild assumptions; (iii) easily incorporates convolutional layers, batch normalization and skip-connections.

As the second contribution, we use this approach to characterize deep neural networks with the most common choices of hyperparameters. We identify the multiplicativity of layer composition as the driving force towards pathologies in feedforward networks: either with the neural network having its signal shrunk into a single point or line; or with the neural network behaving as a noise amplifier with sensitivity

[1] Labatie-AI. Correspondence to: Antoine Labatie <antoine@labatie.ai>.

*Proceedings of the $36^{th}$ International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

exploding with depth. In contrast, we identify the combined action of batch normalization and skip-connections as responsible for bypassing this multiplicativity and relieving from pathologies in batch-normalized residual networks.

*Our results can be fully reproduced with the source code available at* `https://github.com/alabatie/moments-dnns`.

## 2. Propagation

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

**Clean Propagation.** We first consider a random tensorial input $\mathbf{x} \equiv \mathbf{x}^0 \in \mathbb{R}^{n\times\cdots\times n\times N_0}$, 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, fixed spatial extent $n$, and activation function $\phi$.[1] At each layer $l \geq 1$, we denote $N_l$ the number of channels or *width*, $K_l$ the convolutional spatial extent, $\mathbf{x}^l, \mathbf{y}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ the post-activations and pre-activations, $\boldsymbol{\omega}^l \in \mathbb{R}^{K_l\times\cdots\times K_l\times N_{l-1}\times N_l}$ the weights, and $\mathbf{b}^l \in \mathbb{R}^{N_l}$ the biases. Later in our analysis, the model parameters $\boldsymbol{\omega}^l$, $\mathbf{b}^l$ will be considered as random, but for now they are considered as *fixed*. At each layer, the propagation is given by

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

with $*$ the convolution and $\boldsymbol{\beta}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ 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.** To make our setup more realistic, we next suppose that the input signal $\mathbf{x}$ is corrupted by an input noise $\mathrm{d}\mathbf{x} \equiv \mathrm{d}\mathbf{x}^0 \in \mathbb{R}^{n\times\cdots\times n\times N_0}$ having small *iid* components such that $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i \mathrm{d}\mathbf{x}_j] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{ij}$, with $\sigma_{\mathrm{d}\mathbf{x}} \ll 1$ and $\delta_{ij}$ the Kronecker delta for multidimensional indices $i, j$. We denote $\Phi_l(\mathbf{x}) \equiv \mathbf{x}^l$, with $\Phi_l$ the neural network mapping from layer 0 to $l$, and we consider the simultaneous propagation of the signal $\Phi_l(\mathbf{x})$ and the noise $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$. At each layer, this simultaneous propagation is given at first order by

$$\begin{aligned} \mathbf{y}^l &= \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l, & \mathrm{d}\mathbf{y}^l &= \boldsymbol{\omega}^l * \mathrm{d}\mathbf{x}^{l-1}, & (1) \\ \mathbf{x}^l &= \phi(\mathbf{y}^l), & \mathrm{d}\mathbf{x}^l &= \phi'(\mathbf{y}^l) \odot \mathrm{d}\mathbf{y}^l, & (2) \end{aligned}$$

with $\odot$ the element-wise tensor multiplication. The tensor $\mathrm{d}\mathbf{x}^l$ resulting from the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$ in Eq. (1) and Eq. (2) approximates arbitrarily well the noise $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$ as $\sigma_{\mathrm{d}\mathbf{x}} \to 0$ [C.1]. For simplicity, we will keep the terminology of *noise* when referring to $\mathrm{d}\mathbf{x}^l$.

From Eq. (1) and Eq. (2), we see that $\mathbf{x}^l$, $\mathbf{y}^l$ only depend on the input signal $\mathbf{x}$, and that $\mathrm{d}\mathbf{x}^l$ depends linearly on the input noise $\mathrm{d}\mathbf{x}$ when $\mathbf{x}$ is *fixed*. As a consequence, $\mathrm{d}\mathbf{x}^l$ stays centered with respect to $\mathrm{d}\mathbf{x}$ such that $\forall \mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}] = 0$, where from now on the spatial position is denoted as $\alpha$ and the channel as c.

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

Some results of our analysis will apply for any choice of $\phi$, but unless otherwise stated, we restrict to the most common choice: $\phi(\cdot) \equiv \mathrm{ReLU}(\cdot) = \max(\cdot, 0)$. Even though ReLU is not differentiable at 0, we still define $\mathrm{d}\mathbf{x}^l$ as the result of the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$ in Eq. (1) and Eq. (2) with the convention $\phi'(0) \equiv 1/2$ [C.2].

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

## 3. Data Randomness

Now we may turn our attention to the data distributions of signal and noise: $P_{\mathbf{x},\boldsymbol{\alpha}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}(\mathrm{d}\mathbf{x}^l)$. To outline the importance of these distributions, the output of an $L$-layer neural network can be expressed by layer composition as $(\mathbf{x}^L, \mathrm{d}\mathbf{x}^L) = \tilde{\Phi}_{l,L}(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$, with $\tilde{\Phi}_{l,L}$ the mapping of the signal and noise by the *upper neural network* from layer $l < L$ to layer $L$. The upper neural network thus receives $\mathbf{x}^l$ as input signal and $\mathrm{d}\mathbf{x}^l$ as input noise, implying that it can only have a chance to do any better than random guessing when: (i) $\mathbf{x}^l$ is meaningful; (ii) $\mathrm{d}\mathbf{x}^l$ is under control. Namely, when $P_{\mathbf{x},\boldsymbol{\alpha}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}(\mathrm{d}\mathbf{x}^l)$ are not affected by pathologies. We will make this argument as well as the notion of *pathology* more precise in Section 3.2 after a few prerequisite definitions.

### 3.1. Characterizing Data Distributions

Using $\mathbf{v}^l$ as a placeholder for any tensor of layer $l$ in the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$ – e.g. $\mathbf{y}^l$, $\mathbf{x}^l$, $\mathrm{d}\mathbf{y}^l$, $\mathrm{d}\mathbf{x}^l$ in Eq. (1) and Eq. (2) – we define:

– The *feature map vector* and *centered feature map vector*,

$$\varphi(\mathbf{v}^l, \boldsymbol{\alpha}) \equiv \mathbf{v}^l_{\boldsymbol{\alpha},:}, \qquad \hat{\varphi}(\mathbf{v}^l, \boldsymbol{\alpha}) \equiv \mathbf{v}^l_{\boldsymbol{\alpha},:} - \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\mathbf{v}^l_{\boldsymbol{\alpha},:}],$$[3]

[1] It is possible to relax the assumptions of periodic boundary conditions and constant spatial extent $n$ [B.5]. These assumptions, as well as the assumption of constant width $N_l$ in Section 7, are only made for simplicity of the analysis.

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

[3] Slightly abusively, the notation $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}, \mathbf{v}^l$ is overloaded in the expectation.

with $\mathbf{v}^l_{\boldsymbol{\alpha},:}$ the vectorial slice of $\mathbf{v}^l$ at spatial position $\boldsymbol{\alpha}$. Note that $\varphi(\mathbf{v}^l,\boldsymbol{\alpha})$, $\hat{\varphi}(\mathbf{v}^l,\boldsymbol{\alpha})$ aggregate both the randomness from $(\mathbf{x},\mathrm{d}\mathbf{x})$ which determines the propagation up to $\mathbf{v}^l$, and the randomness from $\boldsymbol{\alpha}$ which determines the spatial position in $\mathbf{v}^l$. These random vectors will enable us to circumvent the tensorial structure of $\mathbf{v}^l$.

– The *non-central moment* and *central moment* of order $p$ for given channel c and averaged over channels,

$$\nu_{p,\mathrm{c}}(\mathbf{v}^l) \equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{v}^l,\boldsymbol{\alpha})_{\mathrm{c}}^p\big], \quad \nu_p(\mathbf{v}^l) \equiv \mathbb{E}_{\mathrm{c}}\big[\nu_{p,\mathrm{c}}(\mathbf{v}^l)\big],$$
$$\mu_{p,\mathrm{c}}(\mathbf{v}^l) \equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\hat{\varphi}(\mathbf{v}^l,\boldsymbol{\alpha})_{\mathrm{c}}^p\big], \quad \mu_p(\mathbf{v}^l) \equiv \mathbb{E}_{\mathrm{c}}\big[\mu_{p,\mathrm{c}}(\mathbf{v}^l)\big].$$

In the particular case of the noise $\mathrm{d}\mathbf{x}^l$, centered with respect to $\mathrm{d}\mathbf{x}$, feature map vectors and centered feature map vectors coincide: $\varphi(\mathrm{d}\mathbf{x}^l,\boldsymbol{\alpha}) = \hat{\varphi}(\mathrm{d}\mathbf{x}^l,\boldsymbol{\alpha})$, such that non-central moments and central moments also coincide: $\nu_{p,\mathrm{c}}(\mathrm{d}\mathbf{x}^l) = \mu_{p,\mathrm{c}}(\mathrm{d}\mathbf{x}^l)$ and $\nu_p(\mathrm{d}\mathbf{x}^l) = \mu_p(\mathrm{d}\mathbf{x}^l)$.

– The *effective rank* (Vershynin, 2010),

$$r_{\text{eff}}(\mathbf{v}^l) \equiv \frac{\operatorname{Tr} \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{v}^l,\boldsymbol{\alpha})\big]}{||\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{v}^l,\boldsymbol{\alpha})\big]||},$$

with $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}$ the covariance matrix and $||\cdot||$ the spectral norm. If we further denote $(\lambda_i)$ the eigenvalues of $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\varphi(\mathbf{v}^l,\boldsymbol{\alpha})]$, then $r_{\text{eff}}(\mathbf{v}^l) = \sum_i \lambda_i / \max_i \lambda_i \geq 1$. Intuitively, $r_{\text{eff}}(\mathbf{v}^l)$ measures the number of effective directions which concentrate the variance of $\varphi(\mathbf{v}^l,\boldsymbol{\alpha})$.

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

$$\chi^l \equiv \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathbf{x}^l)}\right)^{\frac{1}{2}} \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\mu_2(\mathbf{x}^0)}\right)^{-\frac{1}{2}}. \tag{3}$$

To grasp the definition of $\chi^l$, we may consider the signal-to-noise ratio $\mathrm{SNR}^l$ and the noise factor $F^l$,

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

We obtain $\mathrm{SNR}^l_{\text{dB}} = \mathrm{SNR}^0_{\text{dB}} - 20\log_{10}\chi^l$ in logarithmic decibel scale, i.e. that $\chi^l$ 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. Neural networks with $\chi^l > 1$ are noise amplifiers, while neural networks with $\chi^l < 1$ are noise reducers.

Now, to justify our choice of terminology, let us reason in the case where $\mathbf{x}^l = \Phi_l(\mathbf{x}^0)$ is the output signal at the final layer. Then: (i) the variance $\mu_2(\mathbf{x}^l)$ is typically constrained by the task (e.g. binary classification constrains $\mu_2(\mathbf{x}^l)$ to be roughly equal to 1); (ii) the constant rescaling $\Psi_l(\mathbf{x}^0) = \sqrt{\mu_2(\mathbf{x}^l)}/\sqrt{\mu_2(\mathbf{x}^0)} \cdot \mathbf{x}^0$ leads to the same constrained variance: $\mu_2(\Psi_l(\mathbf{x}^0)) = \mu_2(\Phi_l(\mathbf{x}^0))$. The normalized sensitivity $\chi^l$ exactly measures the excess root mean square sensitivity of the neural network mapping $\Phi_l$ relative to the constant rescaling $\Psi_l$ [C.3]. This property is illustrated in Fig. 1.

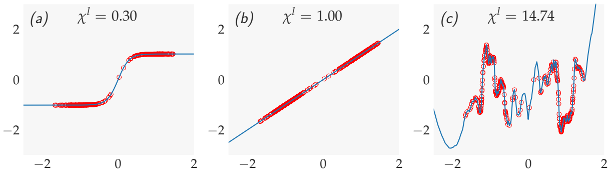

Figure 1: *Illustration of* $\chi^l$ in the fully-connected case with one-dimensional input and output, $N_0 = 1$, $N_l = 1$. We show the full input-output mapping $\Phi_l$ *(blue curves)* and randomly sampled input-output data points $\big(\mathbf{x}^0, \Phi_l(\mathbf{x}^0)\big)$ *(red circles)* for three different neural networks sharing the same input signal $\mathbf{x}^0$ and the same variance in their output signal $\mu_2(\Phi_l(\mathbf{x}^0))$. *(a)* Since input data points $\mathbf{x}^0$ appear in flat regions of $\Phi_l$, the sensitivity is low: $\chi^l < 1$. *(b)* $\Phi_l$ is a constant rescaling: $\chi^l = 1$. *(c)* Since $\Phi_l$ is highly chaotic, the sensitivity is high: $\chi^l > 1$.

As outlined, $\chi^l$ measures the sensitivity to signal perturbation, which is known for being connected to generalization (Rifai et al., 2011; Arpit et al., 2017; Sokolic et al., 2017; Arora et al., 2018; Morcos et al., 2018; Novak et al., 2018; Philipp & Carbonell, 2018). A tightly connected notion is the sensitivity to weight perturbation, also known for being connected to generalization (Hochreiter & Schmidhuber, 1997; Langford & Caruana, 2002; Keskar et al., 2017; Chaudhari et al., 2017; Smith & Le, 2018; Dziugaite & Roy, 2017; Neyshabur et al., 2017; 2018; Li et al., 2018). The connection is seen by noting the equivalence between a noise $\mathrm{d}\boldsymbol{\omega}^l$ on the weights and 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).

### 3.2. Characterizing Pathologies

We are now able to characterize the pathologies, with ill-behaved data distributions, $P_{\mathbf{x},\boldsymbol{\alpha}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}(\mathrm{d}\mathbf{x}^l)$, 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:

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

The pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \to 0$ implies $\mu_2(\tilde{\mathbf{x}}^l) \to 0$, meaning that $\varphi(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})$ becomes point-like concentrated at the point $\tilde{\boldsymbol{\nu}}^l$ of unit $L^2$ norm: $||\tilde{\boldsymbol{\nu}}^l||_2 = 1$ [C.4]. 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 the variance of $\varphi(\mathbf{x}^l, \boldsymbol{\alpha})$ becomes concentrated in a single direction, meaning 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*: $\chi^l \geq \exp(\gamma l) \xrightarrow{l\to\infty} \infty$ for some $\gamma > 0$. Given the noise factor equivalence of Eq. (4), the pathology $\chi^l \to \infty$ implies $\text{SNR}^l \to 0$, meaning that the clean signal $\mathbf{x}^l$ becomes drowned in the noise $\mathrm{d}\mathbf{x}^l$. In the limit of strictly zero signal-to-noise ratio, the upper neural network from layer $l$ to $L$ is limited to random guessing since it only "sees" noise.

## 4. Model Parameters Randomness

We now introduce model parameters as the second source of randomness. We consider networks at initialization, which we suppose is *standard* following He et al. (2015): (i) weights are initialized with $\boldsymbol{\omega}^l \sim \mathcal{N}\big(0, 2\,/\,(K_l^d N_{l-1})\,\boldsymbol{I}\big)$, biases are initialized with zeros; (ii) when pre-activations are batch-normalized, scale and shift batch normalization parameters are initialized with ones and zeros respectively.

Considering networks at initialization is justified in two respects. As the first justification, in the context of Bayesian neural networks, the distribution on model parameters at initialization induces a distribution on input-output mappings which can be seen as the prior encoded by the choice of hyperparameters (Neal, 1996; Williams, 1997).

As the second justification, even in the standard context of non-Bayesian neural networks, it is likely that pathologies at initialization penalize training by hindering optimization. Let us illustrate this argument in two cases:

– In the case of zero-dimensional signal, 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. This case – further associated with vanishing gradients for bounded $\phi$ (Schoenholz et al., 2017) – is known as the "ordered phase" with unit correlation between different inputs, resulting in untrainability (Schoenholz et al., 2017; Xiao et al., 2018).

– In the case of exploding sensitivity, the upper neural network from layer $l$ to $L$ only "sees" noise and its backpropagated gradient is purely noise. Gradient descent then performs random steps and training loss is not decreased. This case – further associated with exploding gradients for batch-normalized $\phi = \text{ReLU}$ or bounded $\phi$ (Schoenholz et al., 2017) – is known as the "chaotic phase" with decorrelation between different inputs, also resulting in untrainability (Schoenholz et al., 2017; Yang & Schoenholz, 2017; Xiao et al., 2018; Philipp & Carbonell, 2018; Yang et al., 2019).

From now on, our methodology is to consider all moment-related quantities, e.g. $\nu_p(\mathbf{x}^l)$, $\mu_p(\mathbf{x}^l)$, $\mu_p(\mathrm{d}\mathbf{x}^l)$, $r_{\text{eff}}(\mathbf{x}^l)$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^l)$, $\chi^l$, as random variables which depend on model parameters. We denote the model parameters as $\Theta^l \equiv (\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \dots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ and use $\theta^l$ as shorthand for $\Theta^l|\Theta^{l-1}$. We further denote the geometric increments of $\nu_2(\mathbf{x}^l)$ as $\delta\nu_2(\mathbf{x}^l) \equiv \nu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^{l-1})$.

**Evolution with Depth.** The evolution with depth of $\nu_2(\mathbf{x}^l)$ can be written as

$$\log\left(\frac{\nu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^0)}\right) = \sum_{k\leq l} \underbrace{\log\delta\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log\delta\nu_2(\mathbf{x}^k)]}_{\underline{s}[\nu_2(\mathbf{x}^k)]} +$$
$$\underbrace{\mathbb{E}_{\theta^k}[\log\delta\nu_2(\mathbf{x}^k)] - \log\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]}_{\underline{m}[\nu_2(\mathbf{x}^k)]} + \underbrace{\log\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]}_{\overline{m}[\nu_2(\mathbf{x}^k)]},$$

where we used $\log\left(\frac{\nu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^0)}\right) = \log\nu_2(\mathbf{x}^l) - \log\nu_2(\mathbf{x}^0) = \sum_{k\leq l}\log\delta\nu_2(\mathbf{x}^k)$ and expressed $\log\delta\nu_2(\mathbf{x}^k)$ with telescoping terms. Denoting $\underline{\delta}\nu_2(\mathbf{x}^k) \equiv \delta\nu_2(\mathbf{x}^k)/\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$ the multiplicatively centered increments of $\nu_2(\mathbf{x}^k)$, we get [C.5]

$$\overline{m}[\nu_2(\mathbf{x}^k)] = \log\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)], \tag{5}$$
$$\underline{m}[\nu_2(\mathbf{x}^k)] = \mathbb{E}_{\theta^k}[\log\underline{\delta}\nu_2(\mathbf{x}^k)], \tag{6}$$
$$\underline{s}[\nu_2(\mathbf{x}^k)] = \log\underline{\delta}\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log\underline{\delta}\nu_2(\mathbf{x}^k)]. \tag{7}$$

**Discussion.** We directly note that: (i) $\overline{m}[\nu_2(\mathbf{x}^k)]$ and $\underline{m}[\nu_2(\mathbf{x}^k)]$ are random variables which depend on $\Theta^{k-1}$, while $\underline{s}[\nu_2(\mathbf{x}^k)]$ is a random variable which depends on $\Theta^k$; (ii) $\underline{m}[\nu_2(\mathbf{x}^k)] < 0$ by log-concavity; (iii) $\underline{s}[\nu_2(\mathbf{x}^k)]$ is centered with $\mathbb{E}_{\theta^k}[\underline{s}[\nu_2(\mathbf{x}^k)]] = 0$ and $\mathbb{E}_{\Theta^k}[\underline{s}[\nu_2(\mathbf{x}^k)]] = 0$.

We further note that each channel provides an independent contribution to $\nu_2(\mathbf{x}^k) = \frac{1}{N_k}\sum_{\mathrm{c}}\nu_{2,\mathrm{c}}(\mathbf{x}^k)$, implying for large $N_k$ that $\underline{\delta}\nu_2(\mathbf{x}^k)$ has low expected deviation to 1 and that $|\log\underline{\delta}\nu_2(\mathbf{x}^k)| \ll 1$, $|\underline{m}[\nu_2(\mathbf{x}^k)]| \ll 1$, $|\underline{s}[\nu_2(\mathbf{x}^k)]| \ll 1$ with high probability. The term $\overline{m}[\nu_2(\mathbf{x}^k)]$ is thus dominating as long as it is not vanishing. The same reasoning applies to other positive moments, e.g. $\mu_2(\mathbf{x}^l)$, $\mu_2(\mathrm{d}\mathbf{x}^l)$.

**Further Notation.** From now on, the geometric increment of any quantity is denoted with $\delta$. The definitions of $\overline{m}$, $\underline{m}$ and $\underline{s}$ in Eq. (5), (6) and (7) are extended to other positive moments of signal and noise, as well as $\chi^l$ with

$$\overline{m}[\chi^l] \equiv \tfrac{1}{2}\big(\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)] - \overline{m}[\mu_2(\mathbf{x}^l)]\big),$$
$$\underline{m}[\chi^l] \equiv \tfrac{1}{2}\big(\underline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)] - \underline{m}[\mu_2(\mathbf{x}^l)]\big),$$
$$\underline{s}[\chi^l] \equiv \tfrac{1}{2}\big(\underline{s}[\mu_2(\mathrm{d}\mathbf{x}^l)] - \underline{s}[\mu_2(\mathbf{x}^l)]\big).$$

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

$$\delta\chi^l = \exp\big(\overline{m}[\chi^l] + \underline{m}[\chi^l] + \underline{s}[\chi^l]\big) \simeq \exp\big(\overline{m}[\chi^l]\big).$$

We stress the *layer-wise* character of this approximation, whose validity only requires $N_l \gg 1$, independently of the depth $l$. This contrasts with the *aggregated* character (up to layer $l$) of the mean field approximation of $\mathbf{y}^l$ as a Gaussian process, whose validity requires not only $N_l \gg 1$ but also – as we will see – that the depth $l$ remains sufficiently small with respect to $N_l$.

## 5. Vanilla Nets

We are fully equipped to characterize deep neural networks at initialization. We start by analyzing vanilla nets which correspond to the propagation introduced in Section 2.

**Theorem 1** (moments of vanilla nets). [D.3] *There exist small constants* $1 \gg m_{\text{min}}, m_{\text{max}}, v_{\text{min}}, v_{\text{max}} > 0$*, random variables* $m_l, m'_l$*,* $s_l, s'_l$ *and events* $A_l$*,* $A'_l$ *of probabilities equal to* $\prod_{k=1}^{l}(1-2^{-N_k})$ *such that*

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

**Discussion.** The conditionality on $A_l$, $A'_l$ is necessary to exclude the collapse: $\nu_2(\mathbf{x}^l) = 0$, $\mu_2(\mathrm{d}\mathbf{x}^l) = 0$, with undefined $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$, occurring e.g. when all elements of $\boldsymbol{\omega}^l$ are strictly negative (Lu et al., 2018). In practice, this conditionality is highly negligible since the probabilities of the complementary events $A_l^{\mathrm{c}}$, $A_l'^{\mathrm{c}}$ decay exponentially in the width $N_l$ [D.4].

Now let us look at the evolution of $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$ under $A_l$, $A'_l$. The initialization He et al. (2015) enforces $\mathbb{E}_{\theta^l}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^{l-1})$ and $\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)] = \mu_2(\mathrm{d}\mathbf{x}^{l-1})$ such that: (i) $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$, $\mathbb{E}_{\Theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]$ are kept stable during propagation; (ii) $\overline{m}[\nu_2(\mathbf{x}^l)]$, $\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)]$ vanish and $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$ are subject to a slow diffusion with small negative drift terms: $\underline{m}[\nu_2(\mathbf{x}^l)] < 0$, $\underline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)] < 0$, and small diffusion terms: $\underline{s}[\nu_2(\mathbf{x}^l)]$, $\underline{s}[\mu_2(\mathrm{d}\mathbf{x}^l)]$ [D.5].[4] The diffusion happens in log-space since layer composition amounts to a multiplicative random effect in real space. It is a finite-width effect since the terms $\underline{m}[\nu_2(\mathbf{x}^l)]$, $\underline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)]$, $\underline{s}[\nu_2(\mathbf{x}^l)]$, $\underline{s}[\mu_2(\mathrm{d}\mathbf{x}^l)]$ also vanish for infinite width.

Fig. 2 illustrates the slowly decreasing negative expectation and slowly increasing variance of $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$, caused by the small negative drift and diffusion terms. Fig. 2 also indicates that $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$ are nearly Gaussian, implying that $\nu_2(\mathbf{x}^l)$, $\mu_2(\mathrm{d}\mathbf{x}^l)$ are nearly lognormal. Two important insights are then provided by the expressions of the expectation: $\exp(\mu + \sigma^2/2)$ and the kurtosis: $\exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 3$ of a lognormal variable $\exp(X)$ with $X \sim \mathcal{N}(\mu, \sigma^2)$. Firstly, the decreasing negative expectation and increasing variance of $\log \nu_2(\mathbf{x}^l)$, $\log \mu_2(\mathrm{d}\mathbf{x}^l)$ act as opposing forces in order to ensure the stabilization of $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]$, $\mathbb{E}_{\Theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]$. Secondly, $\nu_2(\mathbf{x}^l)$, $\mu_2(\mathrm{d}\mathbf{x}^l)$ are stabilized only in terms of expectation and they become fat-tailed distributed as $l \to \infty$.

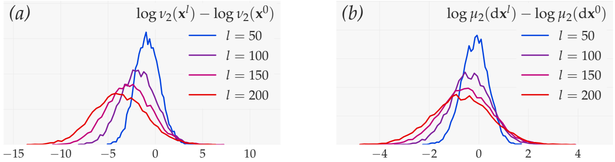

Figure 2: *Slowly diffusing moments of vanilla nets* with $L = 200$ layers of width $N_l = 128$. *(a)* Distribution of $\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)$ for $l = 50, 100, 150, 200$. *(b)* Same for $\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)$.

**Theorem 2** (normalized sensitivity increments of vanilla nets). [D.6] *Denoting* $\mathbf{y}^{l,\pm} \equiv \max\big(\pm\mathbf{y}^l, 0\big)$*, the dominating term under* $\{\mu_2(\mathbf{x}^{l-1}) > 0\}$ *in the evolution of* $\chi^l$ *is*

$$\delta\chi^l \simeq \underbrace{\left(1 - \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,-}\big)}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-\frac{1}{2}}}_{\in[1,\sqrt{2}]}. \qquad (8)$$

**Discussion.** A first consequence is that $\chi^l$ always increases with depth. Another consequence is that only two possibilities of evolution which both lead to pathologies are allowed:

– If sensitivity is exploding: $\chi^l \geq \exp(\gamma l) \to \infty$ with exponential drift $\gamma$ stronger than the slow diffusion of Theorem 1 and if $\nu_2(\mathbf{x}^l)$, $\mu_2(\mathrm{d}\mathbf{x}^l)$ are lognormally distributed as supported by Fig. 2, then Theorem 1 implies the a.s. convergence to the pathology of zero-dimensional signal: $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \to 0$ [D.7].

– Otherwise, geometric increments $\delta\chi^l$ are strongly limited. In the limit $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \to 1$, if the moments of $\tilde{\mathbf{x}}^l \equiv \mathbf{x}^l/\sqrt{\mu_2(\mathbf{x}^l)}$ remain bounded, then Theorem 2 implies the convergence to the pathology of one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$ [D.8] and the convergence to pseudo-linearity, with each additional layer $l$ becoming arbitrarily well approximated by a linear mapping [D.9].

**Experimental Verification.** The evolution with depth of vanilla nets is shown in Fig. 3. From the two possibilities, we observe the case with limited geometric increments: $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \to 1$, the convergence to the pathology of one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$, and the convergence to pseudo-linearity.

[4] Any deviation from He et al. (2015) leads, on the other hand, to pathologies orthogonal to the pathologies of Section 3.2, with either exploding or vanishing constant scalings of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$.

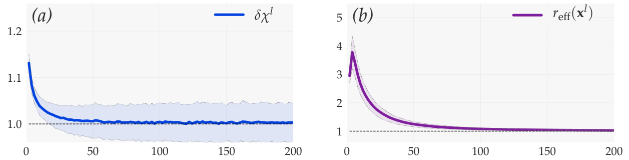

Figure 3: *Pathology of one-dimensional signal for vanilla nets* with $L = 200$ layers of width $N_l = 512$. *(a)* $\delta\chi^l$ such that $\delta\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) \to 1$. *(b)* $r_{\text{eff}}(\mathbf{x}^l)$ indicates one-dimensional signal pathology: $r_{\text{eff}}(\mathbf{x}^l) \to 1$.

The only way that the neural network can achieve pseudo-linearity is by having each one of its $\mathrm{ReLU}$ units either always active or always inactive, i.e. behaving either as zero or as the identity. Our analysis offers theoretical insight into this coactivation phenomenon, previously observed experimentally (Balduzzi et al., 2017; Philipp et al., 2018).

## 6. Batch-Normalized Feedforward Nets

Next we incorporate batch normalization (Ioffe & Szegedy, 2015), which we denote as BN. 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

$$\mathbf{y}^l = \boldsymbol{\omega}^l * \mathbf{x}^{l-1} + \boldsymbol{\beta}^l, \qquad \mathrm{d}\mathbf{y}^l = \boldsymbol{\omega}^l * \mathrm{d}\mathbf{x}^{l-1}, \tag{9}$$

$$\mathbf{z}^l = \mathrm{BN}(\mathbf{y}^l), \qquad \mathrm{d}\mathbf{z}^l = \mathrm{BN}'(\mathbf{y}^l) \odot \mathrm{d}\mathbf{y}^l, \tag{10}$$

$$\mathbf{x}^l = \phi(\mathbf{z}^l), \qquad \mathrm{d}\mathbf{x}^l = \phi'(\mathbf{z}^l) \odot \mathrm{d}\mathbf{z}^l. \tag{11}$$

**Theorem 3** (normalized sensitivity increments of batch-normalized feedforward nets)**.** [E.1] *The dominating term in the evolution of* $\chi^l$ *can be decomposed as*

$$\delta\chi^l = \delta_{\text{BN}}\chi^l \cdot \delta_\phi\chi^l \simeq \exp\big(\overline{m}_{\text{BN}}[\chi^l]\big) \cdot \exp\big(\overline{m}_\phi[\chi^l]\big),$$

$$\exp\big(\overline{m}_{\text{BN}}[\chi^l]\big) \equiv \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-\frac{1}{2}} \mathbb{E}_{\text{c},\theta^l}\left[\frac{\mu_{2,\text{c}}(\mathrm{d}\mathbf{y}^l)}{\mu_{2,\text{c}}(\mathbf{y}^l)}\right]^{\frac{1}{2}},$$

$$\exp\big(\overline{m}_\phi[\chi^l]\big) \equiv \underbrace{\Big(1 - 2\mathbb{E}_{\text{c},\theta^l}\big[\nu_{1,\text{c}}(\mathbf{z}^{l,+})\nu_{1,\text{c}}(\mathbf{z}^{l,-})\big]\Big)^{-\frac{1}{2}}}_{\in[1,\sqrt{2}]}.$$

**Effect of Batch Normalization.** The batch normalization term is such that $\exp(\overline{m}_{\text{BN}}[\chi^l]) \simeq \delta_{\text{BN}}\chi^l$, with $\delta_{\text{BN}}\chi^l$ defined as the increment of $\chi^l$ in the convolution and batch normalization steps of Eq. (9) and Eq. (10). The expression of $\exp(\overline{m}_{\text{BN}}[\chi^l])$ holds for any choice of $\phi$.

This term can be understood intuitively by seeing the different channels c in $\mathbf{y}^l$ as $N_l$ random projections of $\mathbf{x}^{l-1}$ and batch normalization as a modulation of the magnitude for each projection. Since batch normalization uses $\sqrt{\mu_{2,\text{c}}(\mathbf{y}^l)}$ 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 $\chi^l$ owing to the noise factor equivalence of Eq. (4).

Now let us look directly at $\exp(\overline{m}_{\text{BN}}[\chi^l])$ in Theorem 3. If we define the event under which the vectorized weights in channel c have $L^2$ norm equal to $r$: $W^{l,\text{c}}_r \equiv \big\{||\mathrm{vec}(\boldsymbol{\omega}^l_{:,:,:,\text{c}})||_2 = r\big\}$, then spherical symmetry implies that variance increments in channel c from $\mathbf{x}^{l-1}$ to $\mathbf{y}^l$ and from $\mathrm{d}\mathbf{x}^{l-1}$ to $\mathrm{d}\mathbf{y}^l$ have equal expectation under $W^{l,\text{c}}_r$:

$$\frac{\mathbb{E}_{\theta^l|W^{l,\text{c}}_r}[\mu_{2,\text{c}}(\mathbf{y}^l)]}{\mu_2(\mathbf{x}^{l-1})} = \frac{\mathbb{E}_{\theta^l|W^{l,\text{c}}_r}[\mu_{2,\text{c}}(\mathrm{d}\mathbf{y}^l)]}{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}.$$

On the other hand, the variance of these increments depends on the fluctuation of signal and noise in the random direction generated by $\mathrm{vec}(\boldsymbol{\omega}^l_{:,:,:,\text{c}})/||\mathrm{vec}(\boldsymbol{\omega}^l_{:,:,:,\text{c}})||_2$. This depends on the conditioning of signal and noise, i.e. on the magnitude of $r_{\text{eff}}(\mathbf{x}^{l-1})$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^{l-1})$. If we assume that $\mathrm{d}\mathbf{x}^{l-1}$ is well-conditioned, then $\mu_{2,\text{c}}(\mathrm{d}\mathbf{y}^l)/\mu_2(\mathrm{d}\mathbf{x}^{l-1})$ can be treated as a constant and by convexity of the function $x \mapsto 1/x$:

$$\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-1} \mathbb{E}_{\theta^l|W^{l,\text{c}}_r}\left[\frac{\mu_{2,\text{c}}(\mathrm{d}\mathbf{y}^l)}{\mu_{2,\text{c}}(\mathbf{y}^l)}\right] \gtrsim 1,$$

which in turn implies $\exp(\overline{m}_{\text{BN}}[\chi^l]) \gtrsim 1$. 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 $\mu_{2,\text{c}}(\mathbf{y}^l)$ at the denominator and the impact of the convexity. Thus the smaller $r_{\text{eff}}(\mathbf{x}^{l-1})$ and the larger $\exp(\overline{m}_{\text{BN}}[\chi^l])$. This argument is strictly valid for the first step of the propagation wherein the noise has perfect conditioning, resulting in $\exp(\overline{m}_{\text{BN}}[\chi^1]) \geq 1$ [E.2].

**Effect of the Nonlinearity.** The nonlinearity term is such that $\exp(\overline{m}_\phi[\chi^l]) \simeq \delta_\phi\chi^l$, with $\delta_\phi\chi^l$ defined as the increment of $\chi^l$ in the nonlinearity step of Eq. (11). This term is analogous to the term of Eq. (8) for vanilla nets, except that $\nu_{1,\text{c}}(\mathbf{z}^{l,+})\nu_{1,\text{c}}(\mathbf{z}^{l,-})$ is less likely to vanish than $\nu_{1,\text{c}}(\mathbf{y}^{l,+})\nu_{1,\text{c}}(\mathbf{y}^{l,-})/\mu_2(\mathbf{x}^{l-1})$ in Eq. (8) since batch normalization now keeps the signal centered around zero.

**Experimental Verification.** In Fig. 4, we confirm experimentally the pathology of exploding sensitivity: $\chi^l \geq \exp(\gamma l) \to \infty$ for some $\gamma > 0$. We also confirm that: (i) $\mathrm{d}\mathbf{x}^l$ remains well-conditioned, while $\mathbf{x}^l$ becomes ill-conditioned; (ii) $r_{\text{eff}}(\mathbf{x}^l)$ and $\delta_{\text{BN}}\chi^l$ are inversely correlated.

Interestingly, $\delta_\phi\chi^l$ becomes subdominant with respect to $\delta_{\text{BN}}\chi^l$ at large depth. This stems from the fact that $\mathbf{z}^l$ becomes fat-tailed distributed with respect to $\mathbf{x}$, $\boldsymbol{\alpha}$, 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 $|\exp(\overline{m}_\phi[\chi^l]) - 1|$ and thus of $|\delta_\phi\chi^l - 1|$.

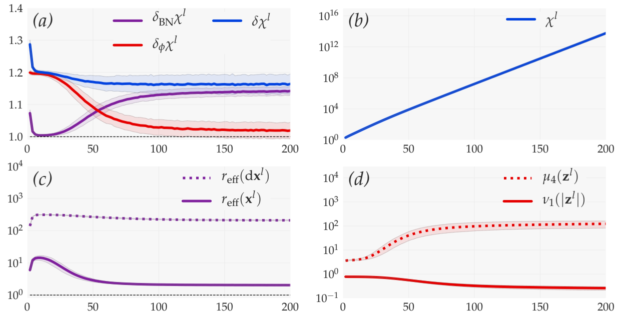

Figure 4: *Pathology of exploding sensitivity for batch-normalized feedforward nets* with $L = 200$ layers of width $N_l = 512$. *(a)* Geometric increments $\delta\chi^l$ decomposed as the product of $\delta_{\text{BN}}\chi^l$ defined as the increment from $(\mathbf{x}^{l-1}, \mathrm{d}\mathbf{x}^{l-1})$ to $(\mathbf{z}^l, \mathrm{d}\mathbf{z}^l)$, and $\delta_\phi\chi^l$ defined as the increment from $(\mathbf{z}^l, \mathrm{d}\mathbf{z}^l)$ to $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$. *(b)* The growth of $\chi^l$ indicates exploding sensitivity pathology: $\chi^l \geq \exp(\gamma l) \to \infty$ for some $\gamma > 0$. *(c)* $\mathbf{x}^l$ becomes ill-conditioned with small $r_{\text{eff}}(\mathbf{x}^l)$. *(d)* $\mathbf{z}^l$ becomes fat-tailed distributed with respect to $\mathbf{x}$, $\boldsymbol{\alpha}$, with large $\mu_4(\mathbf{z}^l)$ and small $\nu_1(|\mathbf{z}^l|)$.

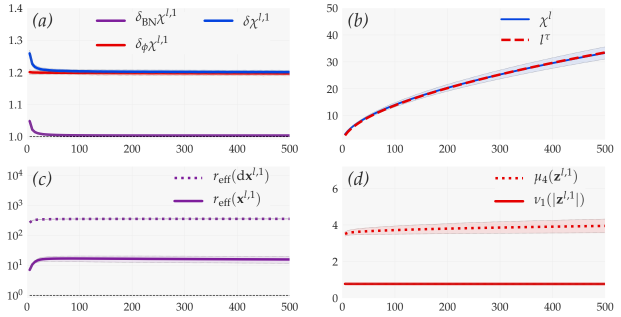

Figure 5: *Well-behaved evolution of batch-normalized resnets* with $L = 500$ residual units comprised of $H = 2$ layers of width $N = 512$. *(a)* Geometric feedforward increments $\delta\chi^{l,1}$ decomposed as the product of $\delta_{\text{BN}}\chi^{l,1}$ defined as the increment from $(\mathbf{y}^{l,0}, \mathrm{d}\mathbf{y}^{l,0})$ to $(\mathbf{z}^{l,1}, \mathrm{d}\mathbf{z}^{l,1})$, and $\delta_\phi\chi^{l,1}$ defined as the increment from $(\mathbf{z}^{l,1}, \mathrm{d}\mathbf{z}^{l,1})$ to $(\mathbf{y}^{l,1}, \mathrm{d}\mathbf{y}^{l,1})$. *(b)* $\chi^l$ has power-law growth. *(c)* $r_{\text{eff}}(\mathbf{x}^{l,1})$ indicates that many directions of signal variance are preserved. *(d)* $\mu_4(\mathbf{z}^{l,1})$, $\nu_1(|\mathbf{z}^{l,1}|)$ indicate that $\mathbf{z}^{l,1}$ has close to Gaussian data distribution.

## 7. Batch-Normalized Resnets

We finish our exploration of deep neural network architectures with the incorporation of skip-connections. From now on, we assume that the width is constant, $N_l = N$, and following He et al. (2016), we adopt the perspective of pre-activation units. The propagation is given by

$$(\mathbf{y}^l, \mathrm{d}\mathbf{y}^l) = (\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1}) + (\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H}), \tag{12}$$

$$\underbrace{\begin{aligned} \mathbf{z}^{l,h} &= \mathrm{BN}(\mathbf{y}^{l,h-1}), & \mathrm{d}\mathbf{z}^{l,h} &= \mathrm{BN}'(\mathbf{y}^{l,h-1}) \odot \mathrm{d}\mathbf{y}^{l,h-1}, \\ \mathbf{x}^{l,h} &= \phi(\mathbf{z}^{l,h}), & \mathrm{d}\mathbf{x}^{l,h} &= \phi'(\mathbf{z}^{l,h}) \odot \mathrm{d}\mathbf{z}^{l,h}, \\ \mathbf{y}^{l,h} &= \boldsymbol{\omega}^{l,h} * \mathbf{x}^{l,h} + \boldsymbol{\beta}^{l,h}, & \mathrm{d}\mathbf{y}^{l,h} &= \boldsymbol{\omega}^{l,h} * \mathrm{d}\mathbf{x}^{l,h}. \end{aligned}}_{\substack{1 \leq h \leq H, \text{ with } H \text{ the number of layers inside residual units} \\ \text{and with } (\mathbf{y}^{l,0}, \mathrm{d}\mathbf{y}^{l,0}) \equiv (\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1})}}$$

If we adopt the convention $(\mathbf{y}^{0,H}, \mathrm{d}\mathbf{y}^{0,H}) \equiv (\mathbf{y}^0, \mathrm{d}\mathbf{y}^0)$, then Eq. (12) can be expanded as

$$(\mathbf{y}^l, \mathrm{d}\mathbf{y}^l) = \sum\nolimits_{k=0}^{l} (\mathbf{y}^{k,H}, \mathrm{d}\mathbf{y}^{k,H}). \tag{13}$$

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

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

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

**Theorem 4** (normalized sensitivity increments of batch-normalized resnets)**.** [F.3] *Suppose that we can bound signal variances:* $\mu_{2,\text{min}} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\text{max}}$ *and feedforward increments:* $\delta_{\text{min}} \lesssim \delta\chi^{l,h} \lesssim \delta_{\text{max}}$ *for all* $l, h$. *Further denote* $\eta_{\text{min}} \equiv \big((\delta_{\text{min}})^{2H}\mu_{2,\text{min}} - \mu_{2,\text{max}}\big)/\mu_{2,\text{max}}$ *and* $\eta_{\text{max}} \equiv \big((\delta_{\text{max}})^{2H}\mu_{2,\text{max}} - \mu_{2,\text{min}}\big)/\mu_{2,\text{min}}$, *as well as* $\tau_{\text{min}} \equiv \frac{1}{2}\eta_{\text{min}}$ *and* $\tau_{\text{max}} \equiv \frac{1}{2}\eta_{\text{max}}$. *Then there exist positive constants* $C_{\text{min}}$, $C_{\text{max}} > 0$ *such that*

$$\left(1 + \frac{\eta_{\text{min}}}{l+1}\right)^{\frac{1}{2}} \lesssim \delta\chi^l \lesssim \left(1 + \frac{\eta_{\text{max}}}{l+1}\right)^{\frac{1}{2}}, \tag{14}$$

$$C_{\text{min}} l^{\tau_{\text{min}}} \lesssim \chi^l \lesssim C_{\text{max}} l^{\tau_{\text{max}}}. \tag{15}$$

**Discussion.** First let us note that Theorem 4 remarkably holds for any choice of $\phi$, with and without batch normalization, as long as the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$, $\delta_{\text{min}}$, $\delta_{\text{max}}$ is ensured. In the case $\phi = \mathrm{ReLU}$, the existence of $\delta_{\text{min}}$, $\delta_{\text{max}}$ is always ensured but the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$ is only ensured when batch normalization controls signal variance inside residual units: $\mu_{2,\text{c}}(\mathbf{z}^{l,H}) = 1$ [F.4].

Now let us get a better grasp of Theorem 4. We see in Eq. (14) that the evolution remains exponential inside residual units since $\eta_{\text{min}}$, $\eta_{\text{max}}$ have an exponential dependence in $H$. However, it is slowed down by the factor $1/(l+1)$ between successive residual units. This stems from the dilution (Philipp et al., 2018) of the residual path $(\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H})$ into the skip-connection path $(\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1})$ with ratio of signal variances: $\mu_2(\mathbf{y}^{l,H})/\big(\mu_2(\mathbf{y}^{l,H}) + \mu_2(\mathbf{y}^{l-1})\big)$ decaying as $1/(l+1)$. If we remove the dilution effect by multiplying the residual branch by 0 (i.e. replacing the scaling in $1/(l+1)$ by a scaling in 1) and if we set $\mu_{2,\text{min}} = \mu_{2,\text{max}}$, then Eq. (14) recovers the feedforward evolution $(\delta_{\text{min}})^H \lesssim \delta\chi^l \lesssim (\delta_{\text{max}})^H$. The dilution is clearly visible in Eq. (13). Namely, each residual unit adds a term $(\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H})$ of increased $\chi^{l,H}$ but its relative contribution to the aggregation gets smaller and smaller with $l$, so that the growth of $\chi^l$ gets slower and slower with $l$.

Since $\frac{1}{2}\log(1+\frac{\eta}{x}) \simeq \frac{\eta}{2x}$ and $\int_{x_0}^{x} \frac{\eta}{2x'}dx' \simeq \log x^{\frac{\eta}{2}}$ for $x \gg 1$, the bounds on $\chi^l = \prod_k \delta\chi^k = \exp\big(\sum_k \log\delta\chi^l\big)$ in Eq. (15) are obtained by integrating the bounds on the logarithm of Eq. (14). A direct consequence of the dilution is thus the power-law evolution of $\chi^l$ instead of the exponential evolution for feedforward nets. Equivalently, when rewriting Eq. (15) as

$$C_{\text{min}} \exp(\tau_{\text{min}} \log l) \lesssim \chi^l \lesssim C_{\text{max}} \exp(\tau_{\text{max}} \log l),$$

the evolution of $\chi^l$ for resnets is equivalent to the evolution of $\chi^{\tau \log l}$ for some $\tau > 0$ for feedforward nets. In other words, the evolution with depth of resnets is the *logarithmic* version of the evolution with depth of feedforward nets.

**Experimental Verification.** The evolution with depth of batch-normalized resnets is shown in Fig. 5. There is a clear parallel between the evolution for $l \leq 500$ in Fig. 5 and the evolution for $l \lesssim 15$ in Fig. 4. This confirms that batch-normalized resnets are slower-to-evolve variants of batch-normalized feedforward nets.

The exponent in the power-law fit of Fig. 5b is notably set to $\tau \equiv \frac{1}{2}(\langle\delta\chi^{l,1}\rangle^{2H} - 1)$, with the feedforward increment $\langle\delta\chi^{l,1}\rangle$ averaged over the whole evolution. This means that Eq. (15) very well describes the evolution of $\chi^l$ in practice.

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 data distribution, as indicated e.g. by $\mu_4(\mathbf{z}^{l,1})$ close to the Gaussian kurtosis of 3. No pathology occurs.

## 8. Discussion and Summary

The novel approach that we introduced for the characterization of deep neural networks at initialization brings three main contributions: (i) it offers a unifying treatment of the broad spectrum of pathologies; (ii) it relies on mild assumptions; (iii) it easily incorporates convolutional layers, batch normalization and skip connections.

Most studies on the convergence of neural networks to Gaussian processes have until now considered the maximal depth $L$ as constant and the width in the limit $N_l \to \infty$ for $l \leq L$. We reversed this perspective by considering the width $N_l$ as large but still bounded and the depth in the limit $l \to \infty$. Then the mean-field approximation of $\mathbf{y}^l$ as a Gaussian process indexed by $\mathbf{x}, \boldsymbol{\alpha}$ eventually becomes invalid:

– In the context of vanilla nets, with e.g. an input $\varphi(\mathbf{x}, \boldsymbol{\alpha})$ constant with respect to $\boldsymbol{\alpha}$ and 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. 2, the $L^2$ norm $||\varphi(\mathbf{x}^l, \boldsymbol{\alpha})||_2^2 = N_l \nu_2(\mathbf{x}^l)$ becomes fat-tailed distributed as $l \to \infty$. For given $\mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$, this means that $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and thus $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}$ become fat-tailed distributed as $l \to \infty$.

– In the context of batch-normalized feedforward nets, with e.g. an input $\varphi(\mathbf{x}, \boldsymbol{\alpha})$ constant with respect to $\boldsymbol{\alpha}$ and uniformly sampled among $M$ points positioned spherically symmetrically in $\mathbb{R}^{N_0}$. Given the evolution of Fig. 4, spherical symmetry together with batch normalization implies that for any given $\mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\Theta^l}[\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}}] = \mathbb{E}_{\Theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^l)] = 0$, $\mathbb{E}_{\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}})^2] = \mathbb{E}_{\Theta^l}[\mu_{2,\mathrm{c}}(\mathbf{z}^l)] = 1$, and $\mathbb{E}_{\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}})^4] = \mathbb{E}_{\Theta^l}[\mu_4(\mathbf{z}^l)] \gg 1$. For given $\mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$, this means that $\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and thus $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}$ become fat-tailed distributed as $l \to \infty$.

Similar observations were made in previous works. Duvenaud et al. (2014) found that the composition of Gaussian processes eventually leads to lognormal and ill-behaved derivatives; Matthews et al. (2018) found that the convergence to Gaussianity as $N_l \to \infty$ becomes slower with respect to $N_l$ as the depth $l$ grows. This stems from the fact that the affine transform at each layer is *additive* with respect to the width dimension, but layer composition is *multiplicative* with respect to the depth dimension. Intuitively, the Central Limit Theorem implies that $\mathbf{y}^l$ becomes normally distributed as $N_l \to \infty$, but lognormally distributed (with fat-tail) as $l \to \infty$.

Beside from this insight, our approach enabled us to characterize deep neural networks with the most common choices of hyperparameters:

– In the case of vanilla nets, the initialization He et al. (2015) limits the evolution of second-order moments of signal and noise. Combined with the limited growth of $\chi^l$, this results in the convergence to the pathology of one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$ and the convergence to neural network pseudo-linearity, with each additional layer $l$ becoming arbitrarily well approximated by a linear mapping.

– In the case of batch-normalized feedforward nets, the pathology of exploding sensitivity: $\chi^l \geq \exp(\gamma l) \to \infty$ for some $\gamma > 0$ has two origins: on the one hand, batch normalization which upweights low-signal pre-activation directions; on the other hand, the nonlinearity $\phi$.

– Finally in the case of resnets, $\chi^l$ 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 $1/(l+1)$. This mechanism is responsible for breaking the circle of depth multiplicativity which causes pathologies for feedforward nets.

## Acknowledgements

Many thanks are due to Jean-Baptiste Fiot for his precious feedback on initial drafts and to the anonymous reviewers for their insightful comments.

## References

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. 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.

Arpit, D., Jastrzebski, S. K., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 233–242, 2017. URL http://proceedings.mlr.press/v70/arpit17a.html.

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W., and McWilliams, B. 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.

Billingsley, P. *Probability and Measure*. Wiley Series in Probability and Statistics. Wiley, 1995. ISBN 9780471007104.

Borovykh, A. A Gaussian Process perspective on Convolutional Neural Networks. *arXiv e-prints*, October 2018. URL https://arxiv.org/abs/1810.10798.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. Entropy-sgd: Biasing gradient descent into wide valleys. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1YfAfcgl.

Chiani, M., Dardari, D., and Simon, M. K. 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.

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *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.

Dziugaite, G. K. and Roy, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*, 2017. URL http://auai.org/uai2017/proceedings/papers/173.pdf.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bklfsi0cKm.

Hanin, B. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 580–589, 2018. URL http://papers.nips.cc/paper/7339-which-neural-net-architectures-give-rise-to-exploding-and-vanishing-gradients.

Hanin, B. and Rolnick, D. How to start training: The effect of initialization and architecture. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pp. 569–579, 2018. URL http://papers.nips.cc/paper/7338-how-to-start-training-the-effect-of-initialization-and-architecture.

He, K., Zhang, X., Ren, S., and Sun, J. 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.

He, K., Zhang, X., Ren, S., and Sun, J. 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, 2016. doi: 10.1007/

978-3-319-46493-0_38. URL https://doi.org/10.1007/978-3-319-46493-0_38.

Hochreiter, S. and Schmidhuber, J. 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.

Ioffe, S. and Szegedy, C. 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.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1oyRlYgg.

Langford, J. and Caruana, R. (not) bounding the true error. In Dietterich, T. G., Becker, S., and Ghahramani, Z. (eds.), *Advances in Neural Information Processing Systems 14*, pp. 809–816. MIT Press, 2002. URL http://papers.nips.cc/paper/1968-not-bounding-the-true-error.pdf.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1EA-M-0Z.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 6391–6401. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/7875-visualizing-the-loss-landscape-of-neural-nets.pdf.

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

Matthews, A. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian Process Behaviour in Wide Deep Neural Networks. *ArXiv e-prints*, April 2018. URL http://adsabs.harvard.edu/abs/2018arXiv180411271M.

Morcos, A. S., Barrett, D. G., Rabinowitz, N. C., and Botvinick, M. On the importance of single directions for generalization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1iuQjxCZ.

Neal, R. M. *Priors for Infinite Networks*, pp. 29–53. Springer New York, New York, NY, 1996. 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.

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring generalization in deep learning. In *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.

Neyshabur, B., Bhojanapalli, S., and Srebro, N. A PAC-bayesian approach to spectrally-normalized margin bounds for neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Skz_WfbCZ.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HJC2SzZCW.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Abolafia, D. A., Pennington, J., and Sohl-dickstein, J. Deep bayesian convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1g30j0qF7.

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

Philipp, G., Song, D., and Carbonell, J. G. Gradients explode - deep networks are shallow - resnet explained. In *International Conference on Learning Representations - Workshop Track*, 2018. URL https://openreview.net/forum?id=HkpYwMZRb.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. 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.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *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.

Rifai, S., Vincent, P., Muller, X., Glorot, X., and Bengio, Y. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 833–840, 2011. URL http://icml-2011.org/papers/455_icmlpaper.pdf.

Roux, N. L. and Bengio, Y. Continuous neural networks. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pp. 404–411. PMLR, 21–24 Mar 2007. URL http://proceedings.mlr.press/v2/leroux07a.html.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=H1W1UN9gg.

Smith, S. L. and Le, Q. V. A bayesian perspective on generalization and stochastic gradient descent. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJij4yg0Z.

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. 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.

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

Williams, C. K. I. Computing with infinite networks. In *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.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *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.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *CoRR*, abs/1902.04760, 2019. URL http://arxiv.org/abs/1902.04760.

Yang, G. and Schoenholz, S. Mean field residual networks: On the edge of chaos. In *Advances in Neural Information Processing Systems 30*, pp. 7103–7114. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/6879-mean-field-residual-networks-on-the-edge-of-chaos.pdf.

Yang, G., Pennington, J., Rao, V., Sohl-Dickstein, J., and Schoenholz, S. S. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyMDXnCcF7.

# A. Details of the Experiments

Fig. 1 considered an input $\mathbf{x}^0$ as a Gaussian mixture, with $\mathbf{x}^0 \sim \mathcal{N}(-1, 0.3^2)$ with probability $1/2$ and $\mathbf{x}^0 \sim \mathcal{N}(1, 0.3^2)$ with probability $1/2$. This input $\mathbf{x}^0$ was propagated into: (a) a single layer with $\phi = \texttt{tanh}$; (b) a single layer with $\phi$ linear; (c) a batch-normalized feedforward net with: $\phi = \mathrm{ReLU}$ and $N_k = 100$ for $1 \leq k < 10$; $\phi$ linear and $N_l = 1$ for $l = 10$.

The experiments of Fig. 2, 3, 4, 5 were made on `cifar10` with a random initial convolution of stride 2 reducing the spatial dimension from 32 to $n = 16$ and increasing the width from 3 to $N_0$. In each case, we considered the convolutional extent $K_l = 3$ and periodic boundary conditions.

In Fig. 2, we considered the width $N_l = 128$ and the total depth $L = 200$. For each realization, we randomly initialized model parameters following He et al. (2015) and randomly sampled $M = 1,024$ images to constitute the input data distribution. For each realization, we then computed the evolution with depth of $\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)$ and $\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)$. The distributions of $\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)$ and $\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)$ shown in Fig. 2 were estimated using $10,000$ such realizations. The limited width – slightly smaller than standard values – had the purpose of limiting computation time in order to gather more realizations.

In Fig. 3, 4, 5, we increased the width to $N_l = 512$. For each realization, we randomly initialized model parameters following He et al. (2015) and randomly sampled $M = 64$ images to constitute the input data distribution. We then computed the evolution with depth of all moment-related quantites. For each quantity, the expectation as well as the $1\sigma$ intervals displayed in Fig. 3, 4, 5 were estimated using $1,000$ such realizations.

Let us make a few remarks:

– The limited number of images $M$ for each experiment enabled to reduce the computation time, in particular penalized by the computation of $r_{\text{eff}}(\mathbf{x}^l)$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^l)$, $r_{\text{eff}}(\mathbf{x}^{l,1})$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^{l,1})$ in Fig. 3, 4, 5. For batch-normalized feedforward nets and batch-normalized resnets, choosing $M$ in the range of standard batch sizes also had the advantage that our setup of batch normalization in *test mode* matched the usual setup of batch normalization in *training mode*.

For vanilla nets in Fig. 2, 3 and batch-normalized resnets in Fig. 5, this reduction of $M$ had very little impact. For batch-normalized feedforward nets in Fig. 4, on the other hand, this reduction of $M$ had the effect of limiting pathologies in the signal. This can be understood by considering $M'$ batch-normalized random points $(\mathbf{z}_0, \dots, \mathbf{z}_{M'})$. In our case, $M'$ is proportional to $M$ but $M' > M$ since the data distribution depends on the input $\mathbf{x}$ *and* the spatial position $\boldsymbol{\alpha}$. By considering the worst-case scenario such that $(\mathbf{z}_0, \dots, \mathbf{z}_{M'}) = (-a, \dots, -a, b, -a, \dots, -a)$:

$$\frac{1}{M'} \sum_i \mathbf{z}_i = \frac{-(M'-1)a + b}{M'}, \quad \frac{1}{M'} \sum_i (\mathbf{z}_i)^2 = \frac{(M'-1)a^2 + b^2}{M'}, \quad \frac{1}{M'} \sum_i (\mathbf{z}_i)^4 = \frac{(M'-1)a^4 + b^4}{M'},$$

$$\frac{1}{M'} \sum_i \mathbf{z}_i = 0, \quad \frac{1}{M'} \sum_i (\mathbf{z}_i)^2 = 1 \implies a = \frac{1}{\sqrt{M'-1}}, \quad b = \sqrt{M'-1}, \quad \frac{1}{M'} \sum_i (\mathbf{z}_i)^4 = \frac{1 + (M'-1)^3}{M'(M'-1)}.$$

This shows that the empirical kurtosis of $(\mathbf{z}_0, \dots, \mathbf{z}_{M'})$ is roughly bounded by $M'$, i.e. that the pathologies of the signal are naturally limited by the number of input images $M$. As a result, for larger $M$ we found that: (i) $r_{\text{eff}}(\mathbf{x}^l)$ gets closer to 1; (ii) $\mu_4(\mathbf{z}^l)$ gets even larger and $\nu_1(|\mathbf{z}^l|)$ gets even smaller; (iii) $|\exp(\overline{m}_{\mathrm{BN}}[\chi^l]) - 1|$ and $|\delta_{\mathrm{BN}}\chi^l - 1|$ get larger; (iv) $|\exp(\overline{m}_{\phi}[\chi^l]) - 1|$ and $|\delta_{\phi}\chi^l - 1|$ get even smaller.

– The dynamics of $|\exp(\overline{m}_{\mathrm{BN}}[\chi^l]) - 1|$ at very low depth in Fig. 4, 5 stems from the input images from `cifar10` having a number of channels equal to $3 \ll N_l = 512$. The signal is therefore ill-conditioned at very low depth and quickly gets better conditioned, implying that $|\exp(\overline{m}_{\mathrm{BN}}[\chi^l]) - 1|$ is non-negligible at very low depth and quickly gets vanishing. This dynamics is brief and occurs before the settling of the main dynamics which leads in particular to the conditioning of the signal degrading again in Fig. 4.

– We tested to set more realistic values for the width $N_l$ in the experiment of Fig. 2. We always observed an absolutely equivalent behaviour apart from the diffusion getting slower with larger $N_l$.

– We tested to change the boundary conditions from periodic to reflective and to zero-padding. We always observed an equivalent behaviour with reflective conditions. As for zero-padding conditions: (i) the evolution of vanilla nets was slightly changed with $r_{\text{eff}}(\mathbf{x}^l)$ converging to a value of roughly 2 instead of 1 due to the creation of new signal directions by zero-padding; (ii) the evolution of batch-normalized feedforward nets and batch-normalized resnets were always equivalent.

– We tested to change the dataset from `cifar10` to `mnist`, with the random initial convolution of stride 2 reducing the spatial dimension from 28 to $n = 14$ and increasing the width from 1 to $N_0$. We observed an equivalent behaviour apart from the signal being slightly more fat-tailed at low depth due to the original images being more fat-tailed in `mnist` than in `cifar10`.

– Finally we tested to change the fuzz parameter $\epsilon$ of batch normalization. The experiments of Fig. 4, 5 used the standard value $\epsilon = 0.001$ but we observed an indistinguishable behaviour when using the value $\epsilon = 0$.

## B. Complementary Definitions and Notations

In this section, we use again $\mathbf{v}^l$ as placeholder for any tensor of layer $l$ in the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$.

### B.1. Receptive Field

**Receptive Field Mapping.** Let us consider the convolution at layer $l$ of an input $\mathbf{v}^{l-1} \in \mathbb{R}^{n\times\cdots\times n\times N_{l-1}}$ from layer $l-1$. The output feature map of the convolution $(\boldsymbol{\omega}^l * \mathbf{v}^{l-1})_{\boldsymbol{\alpha},:}$ 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 from $\mathbf{v}^{l-1}$ of size $K_l^d N_{l-1}$, with $K_l^d$ the spatial extent and $N_{l-1}$ the channel extent. The local input region is called the *receptive field* of $\boldsymbol{\omega}^l * \mathbf{v}^{l-1}$ at spatial position $\boldsymbol{\alpha}$.

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

**Receptive Field Vectors.** The *receptive field vector* and *centered receptive field vector* associated with $\mathbf{v}^{l-1}$ are defined as

$$\rho(\mathbf{v}^{l-1},\boldsymbol{\alpha}) \equiv \mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},:} \quad \text{and} \quad \hat{\rho}(\mathbf{v}^{l-1},\boldsymbol{\alpha}) \equiv \mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},:} - \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},:}],$$

where, slightly abusively, we overloaded the notation $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}, \mathbf{v}^{l-1}$ in the expectation. Again, strictly speaking, $\rho$ and $\hat{\rho}$ depend on $l$ but this is implied by the argument.

### B.2. Propagation with Receptive Field Formulation

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

$$\mathbf{y}^l_{\boldsymbol{\alpha},:} = \boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} + \mathbf{b}^l = \boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} + \boldsymbol{\beta}^l_{\boldsymbol{\alpha},:}, \tag{16}$$

with $\boldsymbol{W}^l \in \mathbb{R}^{N_l\times R_l}$ the suitably reshaped matricial form of $\boldsymbol{\omega}^l$. To lighten notation, we write $\mathbf{y}^l = \boldsymbol{W}^l\mathrm{RF}(\mathbf{x}^{l-1}) + \boldsymbol{\beta}^l$ as a short for the affine transformation of Eq. (16) occuring at all spatial positions $\boldsymbol{\alpha}$. We have the following equivalence between the notations with receptive field and convolution:

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

For vanilla nets, the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$ can be written as

$$\begin{aligned} \mathbf{y}^l &= \boldsymbol{W}^l\mathrm{RF}(\mathbf{x}^{l-1}) + \boldsymbol{\beta}^l, & \mathrm{d}\mathbf{y}^l &= \boldsymbol{W}^l\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1}), \\ \mathbf{x}^l &= \phi(\mathbf{y}^l), & \mathrm{d}\mathbf{x}^l &= \phi'(\mathbf{y}^l)\odot\mathrm{d}\mathbf{y}^l. \end{aligned}$$

For batch-normalized feedforward nets, the simultaneous propagation of $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$ can be written as

$$\begin{aligned} \mathbf{y}^l &= \boldsymbol{W}^l\mathrm{RF}(\mathbf{x}^{l-1}) + \boldsymbol{\beta}^l, & \mathrm{d}\mathbf{y}^l &= \boldsymbol{W}^l\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1}), \\ \mathbf{z}^l &= \mathrm{BN}(\mathbf{y}^l), & \mathrm{d}\mathbf{z}^l &= \mathrm{BN}'(\mathbf{y}^l)\odot\mathrm{d}\mathbf{y}^l, \\ \mathbf{x}^l &= \phi(\mathbf{z}^l), & \mathrm{d}\mathbf{x}^l &= \phi'(\mathbf{z}^l)\odot\mathrm{d}\mathbf{z}^l. \end{aligned}$$

### B.3. Symmetric Propagation

**Symmetric Propagation for Vanilla Nets.** We define additional tensors obtained by *symmetric propagation* at each layer $l$. For vanilla nets, they are given by

$$\begin{aligned} \bar{\mathbf{y}}^l &= -\boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1}) - \boldsymbol{\beta}^l, & \mathrm{d}\bar{\mathbf{y}}^l &= -\boldsymbol{W}^l \mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1}), \\ \bar{\mathbf{x}}^l &= \phi(\bar{\mathbf{y}}^l), & \mathrm{d}\bar{\mathbf{x}}^l &= \phi'(\bar{\mathbf{y}}^l) \odot \mathrm{d}\bar{\mathbf{y}}^l. \end{aligned}$$

Under standard initialization, *the tensor moments have the same distribution with respect to $\theta^l$ 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 $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} \bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$, implying that $\forall \boldsymbol{\alpha}, \mathrm{c}$: $(\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2$. Thus $\forall \mathrm{c}$:

$$\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{y}^l). \tag{17}$$

Now let us consider the second-order moments of the noise tensor:

$$(\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\mathrm{d}\bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathrm{d}\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\mathrm{d}\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 \phi'(\bar{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathrm{d}\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 [\phi'(\mathbf{y}_{\boldsymbol{\alpha},\mathrm{c}})^2 + \phi'(\bar{\mathbf{y}}_{\boldsymbol{\alpha},\mathrm{c}})^2] = (\mathrm{d}\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2, \tag{18}$$

where Eq. (18) was obtained using $\mathrm{d}\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}} = -\mathrm{d}\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} = -\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}}$, as well as the convention $\phi'(0) \equiv 1/2$. Since $\mathrm{d}\mathbf{x}^l$, $\mathrm{d}\bar{\mathbf{x}}^l$, $\mathrm{d}\mathbf{y}^l$ are centered, it follows that $\forall \mathrm{c}$:

$$\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\mathrm{d}\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathrm{d}\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\mathrm{d}\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^l) = \mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^l). \tag{19}$$

**Symmetric Propagation for Batch-Normalized Feedforward Nets.** For batch-normalized feedforward nets, the symmetric propagation at each layer $l$ is given by

$$\bar{\mathbf{y}}^l = -\boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1}) - \boldsymbol{\beta}^l, \qquad \mathrm{d}\bar{\mathbf{y}}^l = -\boldsymbol{W}^l \mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1}), \tag{20}$$

$$\bar{\mathbf{z}}^l = \mathrm{BN}(\bar{\mathbf{y}}^l), \qquad \mathrm{d}\bar{\mathbf{z}}^l = \mathrm{BN}'(\bar{\mathbf{y}}^l) \odot \mathrm{d}\bar{\mathbf{y}}^l, \tag{21}$$

$$\bar{\mathbf{x}}^l = \phi(\bar{\mathbf{z}}^l), \qquad \mathrm{d}\bar{\mathbf{x}}^l = \phi'(\bar{\mathbf{z}}^l) \odot \mathrm{d}\bar{\mathbf{z}}^l. \tag{22}$$

BN in Eq. (21) uses the statistics of $\bar{\mathbf{y}}^l$ such that, under standard initialization, *the tensor moments have the same distribution with respect to $\theta^l$ for both propagations*. We then simply have

$$\bar{\mathbf{z}}^l = -\mathbf{z}^l, \qquad \mathrm{d}\bar{\mathbf{z}}^l = -\mathrm{d}\mathbf{z}^l. \tag{24}$$

The same analysis as before gives $\forall \mathrm{c}$:

$$\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \nu_{2,\mathrm{c}}(\mathbf{z}^l), \tag{25}$$

$$\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\mathrm{d}\bar{\mathbf{x}}^l) = \mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{z}^l). \tag{26}$$

### B.4. Gramian and Covariance Matrices

We adopt the standard definition of the *Gramian matrices* of $\varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$:

$$\begin{aligned} \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &\equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha}) \varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha})^T\big], \\ \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &\equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha}) \hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})^T\big], \\ \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &\equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha}) \rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha})^T\big], \\ \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &\equiv \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha}) \hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})^T\big]. \end{aligned}$$

Then, the *covariance matrices* of $\varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$, $\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})$ are defined as

$$\begin{aligned} \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\varphi(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &= \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] = \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\varphi}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})], \\ \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] &= \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})] = \boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\hat{\rho}(\mathbf{v}^{l-1}, \boldsymbol{\alpha})]. \end{aligned}$$

## B.5. Statistics-Preserving Property

**Statistics-Preserving Property.** $\mathrm{RF}$ is *statistics-preserving* with respect to $\mathbf{v}^{l-1}$ if for any channel c and any index $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$, the random variables $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}} = \rho(\mathbf{v}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}}$ and $\mathbf{v}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}} = \varphi(\mathbf{v}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$, which depend on $\mathbf{x}$, $\mathrm{d}\mathbf{x}$, $\boldsymbol{\alpha}$, have the same distribution: $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}} = \rho(\mathbf{v}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \sim_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}} \mathbf{v}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}} = \varphi(\mathbf{v}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$.

First we will prove that RF is statistics-preserving with respect to $\mathbf{x}^{l-1}$, $\mathrm{d}\mathbf{x}^{l-1}$ when convolutions have periodic boundary conditions and the global spatial extent $n$ is constant. Afterwards, we will provide a possible relaxation of these assumptions. The global spatial extent will be denoted as $n_l$ when it is non-constant.

### B.5.1. Case of Periodic Boundary Conditions and Constant Spatial Extent $n_l = n$

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

**Proof.** Fix a channel c in $\mathbf{v}^{l-1}$, an index $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$, and consider the tensors $\mathbf{v}^{l-1}_{:,\mathrm{c}}$, $\mathrm{RF}(\mathbf{v}^{l-1})_{:,i_\mathrm{c}} \in \mathbb{R}^{n\times\cdots\times n}$. The index $i_\mathrm{c}$ corresponds to a given convolution kernel position $\boldsymbol{\kappa} \in \{1,\dots,K_l\}^d$. Under periodic boundary conditions, this fixed kernel position $\boldsymbol{\kappa}$ implies that each position $\boldsymbol{\alpha}$ in $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}$ originates from a different position $\boldsymbol{\alpha}'$ in the tensor $\mathbf{v}^{l-1}_{\boldsymbol{\alpha}',\mathrm{c}}$. 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 $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}} = \mathbf{v}^{l-1}_{f(\boldsymbol{\alpha}),\mathrm{c}} \sim_{\boldsymbol{\alpha}} \mathbf{v}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}$ when $\mathbf{v}^{l-1}$ is deterministic and $\boldsymbol{\alpha}$ is random. In turn, this implies that $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}} \sim_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}} \mathbf{v}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}$, when $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}$ are random. □

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

**Proof.** This follows immediately from Lemma 1. □

**Corollary 3.** *For any channel* c *and $i_\mathrm{c} \in \mathcal{I}^l_\mathrm{c}$, 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 $\rho(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_\mathrm{c}} \sim_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}} \varphi(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})_\mathrm{c}$. Since the cardinality $|\mathcal{I}^l_\mathrm{c}| = K^d_l$ is the same for all channels* c*, it follows that*

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

*Note that this result always holds in the fully-connected case $n_l = 1$, characterized by $\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha}) = \varphi(\mathbf{x}^{l-1},\boldsymbol{\alpha})$, $\rho(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha}) = \varphi(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})$ and $R_l = N_{l-1}$.*

### B.5.2. Case of Large Spatial Extent $n_l \gg K_l$

**Proposition 4.** *If the convolution stride is one (i.e. $n_{l-1} = n_l$) in most layers and the global spatial extent is much larger than the convolutional spatial extent (i.e. $n_l \gg K_l$) in most layers, then, for any boundary conditions,* $\mathrm{RF}$ *is approximately statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^{l-1}$.*

**Proof.** Fix a layer $l-1$ such that $n_{l-1} = n_l$ and $n_l \gg K_l$. Denote $\mathrm{RF}^{(\mathrm{p})}$ the receptive field mapping associated with periodic boundary conditions. Since $n_{l-1} = n_l \gg K_l$ the receptive fields $\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}, \mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ and $\mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}, \mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ do not intersect boundary regions for most $\boldsymbol{\alpha}$, implying for most $\boldsymbol{\alpha}$:

$$
\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = \mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}, \qquad \mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = \mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}.
$$

This implies for any index $i_\mathrm{c}$ that $P_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big] \simeq P_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big]$ and $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big] \simeq P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_\mathrm{c}}\big]$.

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

## C. Details of Section 3 and Section 4

### C.1. Approximation of $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})$ by $\mathrm{d}\mathbf{x}^l$

We use the definitions and notations from Section B in the context of the propagation of Eq. (1) and Eq. (2). We further suppose that a.s. with respect to $\mathbf{x}$: $\exists r>0$ such that $\Phi_l$ is differentiable in the open ball $\mathcal{B}_r(\mathbf{x})$ of radius $r$ at point $\mathbf{x}$ (see Section C.2 for the justification).

We will prove that

$$\frac{\mu_2(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^l)} \to 0 \quad \text{as } \sigma_{\mathrm{d}\mathbf{x}} \to 0 \text{ (with fixed distributions of } \mathbf{x} \text{ and } \mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}}). \tag{27}$$

Due to the 1-Lipschitzness of $\phi=\mathrm{ReLU}$, under periodic boundary conditions, we have that $\forall \mathbf{t},\mathbf{u},\mathbf{v},\mathbf{w}$:

$$\begin{aligned}\Big(\phi\big(\boldsymbol{W}^l_{\mathrm{c},:}\rho(\mathbf{t},\boldsymbol{\alpha})+\boldsymbol{\beta}^l\big)-\phi\big(\boldsymbol{W}^l_{\mathrm{c},:}\rho(\mathbf{u},\boldsymbol{\alpha})+\boldsymbol{\beta}^l\big)\Big)^2 &\leq ||\boldsymbol{W}^l||^2\cdot||\rho(\mathbf{t},\boldsymbol{\alpha})-\rho(\mathbf{u},\boldsymbol{\alpha})||_2^2 \leq ||\boldsymbol{W}^l||^2\cdot||\mathrm{vec}(\mathbf{t}-\mathbf{u})||_2^2,\\ \Big(\phi'(\mathbf{v}_{\boldsymbol{\alpha},\mathrm{c}})\cdot\boldsymbol{W}^l_{\mathrm{c},:}\rho(\mathbf{w},\boldsymbol{\alpha})\Big)^2 &\leq ||\boldsymbol{W}^l||^2\cdot||\rho(\mathbf{w},\boldsymbol{\alpha})||_2^2 \leq ||\boldsymbol{W}^l||^2\cdot||\mathrm{vec}(\mathbf{w})||_2^2,\end{aligned}$$

with $||\boldsymbol{W}^l||$ the spectral norm of $\boldsymbol{W}^l$. It follows that $\forall \mathbf{x},\mathrm{d}\mathbf{x}$:

$$\begin{aligned}||\mathrm{vec}(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x}))||_2^2 &\leq \left(\prod_{k=1}^l n^d N_l||\boldsymbol{W}^l||^2\right)\cdot||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2,\\ ||\mathrm{vec}(\mathrm{d}\mathbf{x}^l)||_2^2 &\leq \left(\prod_{k=1}^l n^d N_l||\boldsymbol{W}^l||^2\right)\cdot||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2.\end{aligned}$$

This gives:

$$\begin{aligned}||\mathrm{vec}(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l)||_2^2 &\leq 2||\mathrm{vec}(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x}))||_2^2+2||\mathrm{vec}(\mathrm{d}\mathbf{x}^l)||_2^2\\ &\leq 4\left(\prod_{k=1}^l n^d N_l||\boldsymbol{W}^l||^2\right)\cdot||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2\\ &\leq C||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2,\end{aligned}$$

with $C=4\cdot\prod_{k=1}^l n^d N_l||\boldsymbol{W}^l||^2$.

The assumption on the differentiability of $\Phi_l$ implies that $\forall\epsilon>0$, $\exists\eta_\epsilon>0$: $\mathbb{P}_\mathbf{x}\big[\Phi_l$ is differentiable in $\mathcal{B}_{\eta_\epsilon}(\mathbf{x})\big]\geq 1-\epsilon$. Markov's inequality applied to $||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2$ further implies that

$$\mathbb{P}_{\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2>\eta_\epsilon\big]=\mathbb{P}_{\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2>\eta_\epsilon^2\big]\leq\frac{n^d N_0\sigma_{\mathrm{d}\mathbf{x}}^2}{\eta_\epsilon^2}.$$

It then follows that $\forall\epsilon>0$, $\exists\eta_\epsilon,\sigma_\epsilon>0$ such that $\forall\sigma_{\mathrm{d}\mathbf{x}}<\sigma_\epsilon$:

$$\mathbb{P}_{\mathbf{x},\mathrm{d}\mathbf{x}}[A_\epsilon]\geq 1-2\epsilon,$$

with $A_\epsilon=\big\{||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2\leq\eta_\epsilon\big\}\cap\big\{\Phi_l \text{ is differentiable in } \mathcal{B}_{\eta_\epsilon}(\mathbf{x})\big\}$.

Denoting $A^{\mathrm{c}}_\epsilon$ the complementary event of $A_\epsilon$, we deduce that $\forall \sigma_{\mathrm{d}\mathbf{x}} < \sigma_\epsilon$:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[\mathbf{1}_{A^{\mathrm{c}}_\epsilon}||\mathrm{vec}\big(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l\big)||_2^2\big] &\leq \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[\mathbf{1}_{A^{\mathrm{c}}_\epsilon}C||\mathrm{vec}(\mathrm{d}\mathbf{x})||_2^2\big] \\
&\leq C\sigma^2_{\mathrm{d}\mathbf{x}}\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[\mathbf{1}_{A^{\mathrm{c}}_\epsilon}||\mathrm{vec}(\mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}})||_2^2\big] \\
&\leq C\sigma^2_{\mathrm{d}\mathbf{x}}\mathbb{P}_{\mathbf{x},\mathrm{d}\mathbf{x}}[A^{\mathrm{c}}_\epsilon]^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}(\mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}})||_2^4\big]^{\frac{1}{2}} \qquad (28)\\
&\leq C\sigma^2_{\mathrm{d}\mathbf{x}}(2\epsilon)^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}(\mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}})||_2^4\big]^{\frac{1}{2}},
\end{aligned}$$

where we used Cauchy-Schwarz inequality in Eq. (28).

Since $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l=0$ under $A_\epsilon$, it follows that $\forall \sigma_{\mathrm{d}\mathbf{x}} < \sigma_\epsilon$:

$$\begin{aligned}
\frac{\mu_2\big(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l\big)}{\mu_2(\mathrm{d}\mathbf{x}^l)} &= \frac{\frac{1}{n^d N_l}\,\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}\big(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l\big)||_2^2\big]}{\mu_2(\mathrm{d}\mathbf{x})\cdot\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]} \qquad (29)\\
&\leq \frac{\frac{1}{n^d N_l}C(2\epsilon)^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x}}\big[||\mathrm{vec}(\mathrm{d}\mathbf{x}/\sigma_{\mathrm{d}\mathbf{x}})||_2^4\big]^{\frac{1}{2}}}{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]} \leq C'\epsilon^{\frac{1}{2}},
\end{aligned}$$

where we used Proposition 5 in Eq. (29) and appropriately defined the constant $C'$.

Let us finally consider $\epsilon'>0$ and $\epsilon$ such that $C'\epsilon^{\frac{1}{2}}=\epsilon'$. Then $\exists\sigma_\epsilon>0$ such that $\forall\sigma_{\mathrm{d}\mathbf{x}}<\sigma_\epsilon$:

$$\frac{\mu_2\big(\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x})-\Phi_l(\mathbf{x})-\mathrm{d}\mathbf{x}^l\big)}{\mu_2(\mathrm{d}\mathbf{x}^l)} \leq \epsilon',$$

which proves Eq. (27).

### C.2. Assumption that $\Phi_l$ is Differentiable a.s. with respect to $\mathbf{x}$

The *sensitivity equivalence* detailed in Section C.3 relies on the assumption that $\Phi_l(\mathbf{x})$ is differentiable *surely* with respect to $\mathbf{x}$. If $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$, this can be relaxed using subdifferentials by noting that moments with respect to $\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}$ are left unchanged when ignoring zero-probability events.

Now let us justify the assumption that $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$ in the context of the propagation of Eq. (1) and Eq. (2). We denote the receptive field vectors $\rho(\mathbf{x}^{k-1},\boldsymbol{\alpha})$ as in Section B, and we denote $\Theta^l\equiv(\boldsymbol{\omega}^1,\boldsymbol{\beta}^1,\dots,\boldsymbol{\omega}^l,\boldsymbol{\beta}^l)$ as in Section 4. We further assume standard initialization.

Let $A\equiv\big\{\exists r>0$ such that $\Phi_l$ is differentiable in the open ball $\mathcal{B}_r(\mathbf{x})$ of radius $r$ at point $\mathbf{x}\big\}$ be an event depending on $\mathbf{x}$, $\Theta^l$, and let $A^{\mathrm{c}}$ be the complementary event. We will prove that $\mathbb{P}_{\mathbf{x}|\Theta^l}[A]=1$ with probability 1 with respect to $\Theta^l$.

For given $\mathbf{x}$ such that $\forall\boldsymbol{\alpha}$: $\mathbf{x}_{\boldsymbol{\alpha},:}\neq 0$, it is easy to see that

$$A^{\mathrm{c}} \implies \exists k\leq l, \exists\boldsymbol{\alpha},\mathrm{c}: \;\rho(\mathbf{x}^{k-1},\boldsymbol{\alpha})\neq 0 \text{ and } \mathbf{x}^k_{\boldsymbol{\alpha},\mathrm{c}}=0.$$

Under standard initialization, this corresponds to a zero-probability event with respect to $\Theta^l$, meaning that $\mathbb{P}_{\Theta^l|\mathbf{x}}[A]=1-\mathbb{P}_{\Theta^l|\mathbf{x}}[A^{\mathrm{c}}]=1$.

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 $\mathbf{x}_{\boldsymbol{\alpha},:}$ has well-defined probability density function):

$$\mathbb{E}_{\Theta^l}\mathbb{P}_{\mathbf{x}|\Theta^l}[A]=\mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x}|\Theta^l}[\mathbf{1}_A]=\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\Theta^l|\mathbf{x}}[\mathbf{1}_A]=\mathbb{E}_{\mathbf{x}}\mathbb{P}_{\Theta^l|\mathbf{x}}[A]=1. \qquad (30)$$

By contradiction, if there would be non-zero probability with respect to $\Theta^l$ that $\mathbb{P}_{\mathbf{x}|\Theta^l}[A]\neq 1$, then Eq. (30) would not hold. Therefore with probability 1 with respect to $\Theta^l$, $\mathbb{P}_{\mathbf{x}|\Theta^l}[A]=1$, implying that with probability 1 with respect to $\Theta^l$, $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$.

### C.3. Property of Normalized Sensitivity

**Proposition 5.** *The noise tensor* $\mathrm{d}\mathbf{x}^l$ *and the vectorized version of 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* $\mathbf{x} = \mathbf{x}^0$*, are related by:* $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}} = \sqrt{\mu_2(\mathrm{d}\mathbf{x}^l)}/\sqrt{\mu_2(\mathrm{d}\mathbf{x})} = \sqrt{\mu_2(\mathrm{d}\mathbf{x}^l)}/\sqrt{\mu_2(\mathrm{d}\mathbf{x}^0)}$.

**Proof.** Due to the definition of $\mathrm{d}\mathbf{x}^l$ as the first-order approximation of $\Phi_l(\mathbf{x}+\mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x})$:

$$\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} = \big\langle \mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}), \mathrm{vec}(\mathrm{d}\mathbf{x})\big\rangle = \big\langle \mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}), \mathrm{vec}(\mathrm{d}\mathbf{x}^0)\big\rangle,$$

with $\langle\,,\rangle$ the standard dot product in $\mathbb{R}^{n^d N_0}$.

Then due to the white noise property: $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i \mathrm{d}\mathbf{x}_j] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x})\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x}^0)\delta_{ij}$, we deduce that

$$\mathbb{E}_{\mathrm{d}\mathbf{x}}\big[(\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\big] = \mu_2(\mathrm{d}\mathbf{x}) \cdot ||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2,$$

$$\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[(\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})^2\big] = \mu_2(\mathrm{d}\mathbf{x}) \cdot \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big],$$

$$\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x})}\right)^{\frac{1}{2}} = \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\right)^{\frac{1}{2}} = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}}. \qquad \square$$

**Proposition 6.** *Denoting the neural network mapping* $\mathbf{x}^l = \Phi_l(\mathbf{x}) = \Phi_l(\mathbf{x}^0)$ *and the constant rescaling* $\Psi_l(\mathbf{x}) = \sqrt{\mu_2(\mathbf{x}^l)}/\sqrt{\mu_2(\mathbf{x}^0)}\cdot\mathbf{x}^0 = \sqrt{\mu_2(\mathbf{x}^l)}/\sqrt{\mu_2(\mathbf{x})}\cdot\mathbf{x}$ *leading to the same signal variance:* $\mu_2(\Psi_l(\mathbf{x})) = \mu_2(\Phi_l(\mathbf{x}))$*, the normalized sensitivity* $\chi^l$ *exactly measures the excess root mean square sensitivty of the neural network mapping* $\Phi_l$ *relative to the constant rescaling* $\Psi_l$*:*

$$\chi^l = \frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\Phi_l(\mathbf{x})_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}}}{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\Psi_l(\mathbf{x})_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}}} = \frac{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}}}{\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\Psi_l(\mathbf{x})_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}}}.$$

**Proof.** This directly follows from: (i) the definition of $\chi^l$; (ii) the result from Proposition 5; (iii) the fact that the constant rescaling $\Psi_l$ has root mean square sensitivitiy equal to $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}\big[||\mathrm{vec}(\nabla_{\mathbf{x}}\Psi_l(\mathbf{x})_{\boldsymbol{\alpha},\mathrm{c}})||_2^2\big]^{\frac{1}{2}} = \sqrt{\mu_2(\mathbf{x}^l)}/\sqrt{\mu_2(\mathbf{x}^0)}$. □

### C.4. Characterizing Pathologies

We consider the following mean vectors and rescaling of the signal:

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

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

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

The pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$ implies $||\boldsymbol{\nu}^l||_2^2/\big(N_l\nu_2(\mathbf{x}^l)\big) \xrightarrow{l\to\infty} 1$, which in turn implies $\mu_2(\mathbf{x}^l)/||\boldsymbol{\nu}^l||_2^2 \xrightarrow{l\to\infty} 0$, i.e. $\mu_2(\tilde{\mathbf{x}}^l) \xrightarrow{l\to\infty} 0$. It follows that $\varphi(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})$ becomes *point-like* concentrated at point $\tilde{\boldsymbol{\nu}}^l$ of unit $L^2$ norm.

### C.5. Derivation of Eq. (5), (6) and (7)

The quantities $\overline{m}[\nu_2(\mathbf{x}^k)]$, $\underline{m}[\nu_2(\mathbf{x}^k)]$ and $\underline{s}[\nu_2(\mathbf{x}^k)]$ are defined as

$$\begin{aligned}
\overline{m}[\nu_2(\mathbf{x}^k)] &\equiv \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)], \\
\underline{m}[\nu_2(\mathbf{x}^k)] &\equiv \mathbb{E}_{\theta^k}[\log \delta\nu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)], \\
\underline{s}[\nu_2(\mathbf{x}^k)] &\equiv \log \delta\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \delta\nu_2(\mathbf{x}^k)].
\end{aligned}$$

Denoting $\underline{\delta}\nu_2(\mathbf{x}^k) \equiv \delta\nu_2(\mathbf{x}^k)/\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$ the multiplicatively centered increments of $\nu_2(\mathbf{x}^k)$, the term $\underline{m}[\nu_2(\mathbf{x}^k)]$ can be expressed as

$$\begin{aligned}
\underline{m}[\nu_2(\mathbf{x}^k)] &= \mathbb{E}_{\theta^k}\Big[\log\Big(\underline{\delta}\nu_2(\mathbf{x}^k)\,\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]\Big)\Big] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] \\
&= \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)] + \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] \qquad (31) \\
&= \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)],
\end{aligned}$$

where we used $\mathbb{E}_{\theta^k}[\log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]] = \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$ in Eq. (31). The term $\underline{s}[\nu_2(\mathbf{x}^k)]$ can be expressed as

$$\begin{aligned}
\underline{s}[\nu_2(\mathbf{x}^k)] &= \log\Big(\underline{\delta}\nu_2(\mathbf{x}^k)\,\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]\Big) - \mathbb{E}_{\theta^k}\Big[\log\Big(\underline{\delta}\nu_2(\mathbf{x}^k)\,\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]\Big)\Big] \\
&= \log \underline{\delta}\nu_2(\mathbf{x}^k) + \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] - \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] \qquad (32) \\
&= \log \underline{\delta}\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)],
\end{aligned}$$

where we used again $\mathbb{E}_{\theta^k}[\log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]] = \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$ in Eq. (32).

## D. Details of Section 5

### D.1. Lemmas on Weak Convergence

**Weak Convergence.** The sequence of random variables $(X_k)_{k\in\mathbb{N}}$ *converges weakly* to the random variable $X$ if $\mathbb{P}[X_k \leq a] \xrightarrow{k\to\infty} \mathbb{P}[X \leq a]$ for every continuity point $a$ of the function $x \mapsto \mathbb{P}[X \leq x]$. We then write $X_k \Rightarrow X$.

**Tightness.** The sequence of random variables $(X_k)_{k\in\mathbb{N}}$ is *tight* if

$$\forall\epsilon, \exists a_\epsilon, b_\epsilon \in \mathbb{R} : \inf_k \mathbb{P}\big[X_k \in ]a_\epsilon, b_\epsilon]\big] \geq 1-\epsilon.$$

**Uniform Integrability.** The sequence of random variables $(X_k)_{k\in\mathbb{N}}$ is *uniformly integrable* if

$$\sup_k \mathbb{E}\big[\mathbf{1}_{\{|X_k|\geq M\}}|X_k|\big] \xrightarrow{M\to\infty} 0.$$

**Lemma 7** (Theorem 25.7 in Billingsley (1995))**.** *Consider a real-valued function $h$, continuous everywhere apart from a finite set of discontinuity points $D_h = \{x_1, \dots, x_p\}$. Then $h$ is measurable and if $X_k \Rightarrow X$ with $\mathbb{P}[X \in D_h] = 0$, then $h(X_k) \Rightarrow h(X)$.*

**Lemma 8** (Theorem 25.10 in Billingsley (1995), known as Prokhorov's theorem)**.** *If the sequence of random variables $(X_k)_{k\in\mathbb{N}}$ is tight, then it admits a weakly convergent subsequence, i.e. there exists a sequence $(i_k)_{k\in\mathbb{N}}$ of strictly increasing indices and a random variable $X$ such that $X_{i_k} \Rightarrow X$.*

**Lemma 9** (Theorem 25.12 in Billingsley (1995))**.** *If the sequence of random variables $(X_k)_{k\in\mathbb{N}}$ is uniformly integrable and if $X_k \Rightarrow X$, then $X$ has well-defined expectation and $\mathbb{E}[X_k] \xrightarrow{k\to\infty} \mathbb{E}[X]$.*

### D.2. Lemma on the Sum of Increments

**Lemma 10.** *Let us consider a sequence $(X_k)_{k\in\mathbb{N}}$ of random variables and a decreasing sequence of events $(A_k)_{k\in\mathbb{N}}$, which both depend on $\Theta^k$. Let us suppose that $\mathbb{P}_{\theta^k|A_{k-1}}[A_k]$ does not depend on $\Theta^{k-1}$ and let us denote under $A_k$:*

$$Y_k \equiv \mathbb{E}_{\theta^k|A_k}[X_k], \qquad Z_k \equiv X_k - \mathbb{E}_{\theta^k|A_k}[X_k].$$

*Let us further suppose that there exist constants $m_{\text{min}}$, $m_{\text{max}}$, $v_{\text{min}}$, $v_{\text{max}}$ such that $\forall k$, under $A_k$:*

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

*Then it follows that*

*(i) The random variables $Z_k$ are centered and non-correlated such that $\forall k$, $\forall k' \neq k$:*

$$\mathbb{E}_{\Theta^k|A_k}[Z_k] = 0, \qquad \mathbb{E}_{\Theta^{\max(k,k')}|A_{\max(k,k')}}[Z_k Z_{k'}] = 0.$$

*(ii) There exist random variables $m_l$ and $s_l$ such that under $A_l$:*

$$\sum_{k=1}^{l} X_k = l m_l + \sqrt{l} s_l, \quad m_{\text{min}} \leq m_l \leq m_{\text{max}}, \quad \mathbb{E}_{\Theta^l|A_l}[s_l] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\text{max}}.$$

**Proof of (i).** First we show that $Z_k$ is centered under $A_k$:

$$\mathbb{E}_{\theta^k|A_k}[Z_k] = \mathbb{E}_{\theta^k|A_k}[X_k] - \mathbb{E}_{\theta^k|A_k}[X_k] = 0, \tag{33}$$

$$\mathbb{E}_{\Theta^k|A_k}[Z_k] = \frac{\mathbb{E}_{\Theta^{k-1}|A_{k-1}}[\mathbb{E}_{\theta^k|A_{k-1}}[\mathbf{1}_{A_k} Z_k]]}{\mathbb{P}_{\Theta^k|A_{k-1}}[A_k]} = \mathbb{E}_{\Theta^{k-1}|A_{k-1}}\left[\frac{\mathbb{E}_{\theta^k|A_{k-1}}[\mathbf{1}_{A_k} Z_k]}{\mathbb{P}_{\theta^k|A_{k-1}}[A_k]}\right] = \mathbb{E}_{\Theta^{k-1}|A_{k-1}}\big[\mathbb{E}_{\theta^k|A_k}[Z_k]\big] = 0.$$

Now for $k < k'$, we have $k \leq k' - 1$ and thus $Z_k$ is fully determined by $\Theta^{k'-1}$. Then we can write

$$\mathbb{E}_{\Theta^{k'}|A_{k'}}[Z_k Z_{k'}] = \mathbb{E}_{\Theta^{k'-1}|A_{k'-1}} \mathbb{E}_{\theta^{k'}|A_{k'}}[Z_k Z_{k'}] = \mathbb{E}_{\Theta^{k'-1}|A_{k'-1}}\big[Z_k \, \mathbb{E}_{\theta^{k'}|A_{k'}}[Z_{k'}]\big] = 0,$$

where we used Eq. (33). □

**Proof of (ii).** First we note that

$$\mathrm{Var}_{\Theta^k|A_k}\big[Z_k\big] = \mathbb{E}_{\Theta^k|A_k}\big[Z_k^2\big] = \mathbb{E}_{\Theta^{k-1}|A_{k-1}} \mathbb{E}_{\theta^k|A_k}\big[Z_k^2\big] = \mathbb{E}_{\Theta^{k-1}|A_{k-1}} \mathrm{Var}_{\theta^k|A_k}\big[Z_k\big].$$

Combined with the hypothesis that $v_{\text{min}} \leq \mathrm{Var}_{\theta^k|A_k}[Z_k] \leq v_{\text{max}}$, we deduce that

$$v_{\text{min}} \leq \mathrm{Var}_{\Theta^k|A_k}\big[Z_k\big] \leq v_{\text{max}}. \tag{34}$$

Now let us denote $M_l \equiv \sum_{k=1}^{l} Y_k$ and $S_l \equiv \sum_{k=1}^{l} Z_k$. Then, using (i), we get that

$$\begin{aligned}
\mathbb{E}_{\Theta^l|A_l}\big[S_l\big] &= \sum\nolimits_k \mathbb{E}_{\Theta^l|A_l}[Z_k] = 0, \\
\mathrm{Var}_{\Theta^l|A_l}\big[S_l\big] = \mathbb{E}_{\Theta^l|A_l}\big[S_l^2\big] &= \sum\nolimits_{k,k'} \mathbb{E}_{\Theta^l|A_l}\big[Z_k Z_{k'}\big] \\
&= \sum\nolimits_k \mathbb{E}_{\Theta^k|A_k}\big[Z_k^2\big] = \sum\nolimits_k \mathrm{Var}_{\Theta^k|A_k}\big[Z_k\big].
\end{aligned} \tag{35}$$

The hypothesis implies under $A_l$ that $lm_{\text{min}} \leq M_l \leq lm_{\text{max}}$, while Eq. (34) and Eq. (35) together imply that $lv_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}[S_l] \leq lv_{\text{max}}$. If we define $m_l \equiv M_l/l$ and $s_l \equiv S_l/\sqrt{l}$, then $\sum_{k=1}^l X_k = \sum_{k=1}^l Y_k + \sum_{k=1}^l Z_k$ can be written as required under $A_l$:

$$\sum_{k=1}^{l} X_k = M_l + S_l = lm_l + \sqrt{l}s_l, \quad m_{\text{min}} \leq m_l \leq m_{\text{max}}, \quad \mathbb{E}_{\Theta^l|A_l}[s_l] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\text{max}}. \qquad \square$$

### D.3. Proof of Theorem 1

**Theorem 1** (moments of vanilla nets)**.** *There exist small constants* $1 \gg m_{\text{min}}, m_{\text{max}}, v_{\text{min}}, v_{\text{max}} > 0$*, random variables* $m_l, m'_l$*,* $s_l, s'_l$ *and events* $A_l$*,* $A'_l$ *of probabilities equal to* $\prod_{k=1}^l (1 - 2^{-N_k+1})$ *such that*

$$\text{Under } A_l\text{:} \quad \log\left(\frac{\nu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^0)}\right) = -lm_l + \sqrt{l}s_l, \quad m_{\text{min}} \leq m_l \leq m_{\text{max}}, \quad \mathbb{E}_{\Theta^l|A_l}\big[s_l\big] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}\big[s_l\big] \leq v_{\text{max}},$$

$$\text{Under } A'_l\text{:} \quad \log\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\right) = -lm'_l + \sqrt{l}s'_l, \quad m_{\text{min}} \leq m'_l \leq m_{\text{max}}, \quad \mathbb{E}_{\Theta^l|A'_l}\big[s'_l\big] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A'_l}\big[s'_l\big] \leq v_{\text{max}}.$$

#### D.3.1. Proof Introduction

Using the definitions and notations from Section B, denoting $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \dots, \lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{G}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$ and denoting $\hat{\boldsymbol{W}}^l \equiv \boldsymbol{W}^l(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$, we get that $\forall$c:

$$\begin{aligned}\nu_{2,\text{c}}(\mathbf{y}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[(\mathbf{y}^l_{\boldsymbol{\alpha},\text{c}})^2\big] = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^l_{\text{c},:}\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\big)^2\Big] \\ &= \sum_i \big(\hat{\boldsymbol{W}}^l_{\text{c},i}\big)^2 \lambda_i = R_l\, \nu_2(\mathbf{x}^{l-1}) \sum_i \big(\hat{\boldsymbol{W}}^l_{\text{c},i}\big)^2 \hat{\lambda}_i,\end{aligned}$$

where we defined $\hat{\lambda}_i \equiv \lambda_i \,/\, \sum_j \lambda_j$ and used $\sum_j \lambda_j = \mathrm{Tr}\, \boldsymbol{G}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\big] = R_l \nu_2(\mathbf{x}^{l-1})$ by Corollary 3.

Let us further define

$$u^l_{\text{c}} \equiv \begin{cases} \frac{\nu_{2,\text{c}}(\mathbf{x}^l)}{\nu_{2,\text{c}}(\mathbf{x}^l) + \nu_{2,\text{c}}(\bar{\mathbf{x}}^l)} & \text{if } \nu_{2,\text{c}}(\mathbf{x}^l) + \nu_{2,\text{c}}(\bar{\mathbf{x}}^l) > 0 \\ \frac{1}{2} & \text{otherwise.} \end{cases}$$

Combined with $\nu_{2,\text{c}}(\mathbf{y}^l) = \nu_{2,\text{c}}(\mathbf{x}^l) + \nu_{2,\text{c}}(\bar{\mathbf{x}}^l)$ by Eq. (17), we get that $\forall$c, under $\big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

$$\nu_{2,\text{c}}(\mathbf{x}^l) + \nu_{2,\text{c}}(\bar{\mathbf{x}}^l) = R_l\, \nu_2(\mathbf{x}^{l-1}) \sum_i \big(\hat{\boldsymbol{W}}^l_{\text{c},i}\big)^2 \hat{\lambda}_i, \tag{36}$$

$$\nu_{2,\text{c}}(\mathbf{x}^l) = u^l_{\text{c}}\, R_l\, \nu_2(\mathbf{x}^{l-1}) \sum_i \big(\hat{\boldsymbol{W}}^l_{\text{c},i}\big)^2 \hat{\lambda}_i. \tag{37}$$

Now combining Eq. (36) with the symmetry of the propagation: $\nu_{2,\text{c}}(\mathbf{x}^l) \sim_{\theta^l} \nu_{2,\text{c}}(\bar{\mathbf{x}}^l)$, and the assumption of standard initialization: $\boldsymbol{W}^l_{\text{c},:} \sim_{\theta^l} \hat{\boldsymbol{W}}^l_{\text{c},:} \sim_{\theta^l} \mathcal{N}(0, 2\,/\,R_l \boldsymbol{I})$, we get that $\forall$c, under $\big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

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

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

$$\mathbb{E}_{\theta^l}[\delta\nu_2(\mathbf{x}^l)] = 1. \tag{38}$$

Next, let us define

$$v^l_\mathrm{c} \equiv \begin{cases} 0 & \text{if } u^l_\mathrm{c} < \frac{1}{2} \\ 1 & \text{if } u^l_\mathrm{c} > \frac{1}{2} \\ \tilde{v}^l_\mathrm{c} & \text{if } u^l_\mathrm{c} = \frac{1}{2} \end{cases},$$

with $\tilde{v}^l_\mathrm{c} \sim \text{Bernouilli}(1/2)$ independent of $\boldsymbol{\omega}^l$ and $\boldsymbol{\beta}^l$.

Conditionally on $u^l_\mathrm{c} = 1/2$: $v^l_\mathrm{c} \sim \text{Bernouilli}(1/2)$, independently of $\nu_{2,\mathrm{c}}(\mathbf{y}^l)$ and $||\boldsymbol{W}^l_{\mathrm{c},:}||_2$. And conditionally on $u^l_\mathrm{c} \neq 1/2$: $v^l_\mathrm{c} \sim \text{Bernouilli}(1/2)$, independently of $\nu_{2,\mathrm{c}}(\mathbf{y}^l)$ and $||\boldsymbol{W}^l_{\mathrm{c},:}||_2$. It follows that $v^l_\mathrm{c} \sim \text{Bernouilli}(1/2)$, independently of $\nu_{2,\mathrm{c}}(\mathbf{y}^l)$ and $||\boldsymbol{W}^l_{\mathrm{c},:}||_2$.

Defining $B_l \equiv \{\exists \mathrm{c} : v^l_\mathrm{c} = 1\}$ we get that $\mathbb{P}_{\theta^l}[B_l] = 1 - 2^{-N_l}$. We also get that $B_l$ is independent of $\big(||\boldsymbol{W}^l_{\mathrm{c},:}||_2\big)_{1 \leq \mathrm{c} \leq N_l}$ and thus of $||\boldsymbol{W}^l||_F$. *This will be useful later in the course of this proof.*

Denoting $A_l = \bigcap_{k=1}^l \Big(B_k \cap \big\{\nu_2(\mathbf{x}^k) \neq 0\big\}\Big)$, we have that

$$\mathbb{P}_{\theta^l|A_{l-1}}\big[A_l\big] = \mathbb{P}_{\theta^l|A_{l-1}}\big[B_l \cap \{\nu_2(\mathbf{x}^l) \neq 0\}\big] = \mathbb{P}_{\theta^l|A_{l-1}}\big[B_l\big] = 1 - 2^{-N_l},$$
$$\mathbb{P}_{\Theta^l}\big[A_l\big] = \mathbb{P}_{\Theta^l}\Big[\bigcap\nolimits_{k=1}^l A_k\Big] = \prod\nolimits_{k=1}^l \mathbb{P}_{\theta^k|A_{k-1}}\big[A_k\big] = \prod\nolimits_{k=1}^l \big(1 - 2^{-N_k}\big).$$

where we used $\mathbb{P}_{\theta^l|A_{l-1}}\big[B_l \cap \big\{\nu_2(\mathbf{x}^l) \neq 0\big\}\big] = \mathbb{P}_{\theta^l|A_{l-1}}\big[B_l\big]$ due to $\mathbb{P}_{\theta^l|B_l \cap A_{l-1}}\big[\nu_2(\mathbf{x}^l) \neq 0\big] = 1$.

Now 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, Eq. (37) implies that $\exists (w_i)_{1 \leq i \leq R_l}$ such that under $B_l \cap \big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

$$(w_i)_{1 \leq i \leq R_l} \sim \mathcal{N}(0, 2\,/\,R_l\boldsymbol{I}), \qquad \frac{1}{N_l}\left(\frac{1}{2}\right) R_l \nu_2(\mathbf{x}^{l-1}) \sum_{i=1}^{R_l} w_i^2 \hat{\lambda}_i \leq \frac{1}{N_l} \sum_{\mathrm{c}=1}^{N_l} \nu_{2,\mathrm{c}}(\mathbf{x}^l),$$
$$(w_i)_{1 \leq i \leq R_l} \sim \mathcal{N}(0, 2\,/\,R_l\boldsymbol{I}), \qquad \frac{R_l}{2N_l} \sum_{i=1}^{R_l} w_i^2 \hat{\lambda}_i \leq \delta\nu_2(\mathbf{x}^l).$$

On the other hand, $\exists (w_{i,j})_{1 \leq i \leq R_l,\, 1 \leq j \leq N_l}$ such that under $B_l \cap \big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

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

Denoting Chi-Squared$(1)$ and Chi-Squared$(N_l R_l)$ the chi-squared distributions with $1$ and $N_l R_l$ degrees of freedom respectively, $\exists w_\text{min}, w_\text{max}$ such that under $B_l \cap \big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

$$w_\text{min} \sim \frac{R_l}{2N_l} \frac{2}{R_l} \frac{1}{R_l} \text{Chi-Squared}(1), \quad w_\text{max} \sim \frac{R_l}{N_l} \frac{2}{R_l} \text{Chi-Squared}(N_l R_l), \quad w_\text{min} \leq \delta\nu_2(\mathbf{x}^l) \leq w_\text{max},$$
$$w_\text{min} \sim \frac{1}{N_l R_l} \text{Chi-Squared}(1), \qquad w_\text{max} \sim \frac{2}{N_l} \text{Chi-Squared}(N_l R_l), \qquad w_\text{min} \leq \delta\nu_2(\mathbf{x}^l) \leq w_\text{max}, \tag{39}$$

where we used $\max_i \hat{\lambda}_i \geq \frac{1}{R_l}$.

Simply replacing $\mathbf{x}^l$ by $\mathrm{d}\mathbf{x}^l$, $\mathbf{y}^l$ by $\mathrm{d}\mathbf{y}^l$, $\boldsymbol{G}_{\mathbf{x},\boldsymbol{\alpha}}$ by $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}$, using Eq. (19) instead of Eq. (17) and the identity with $\mu_2(\mathrm{d}\mathbf{x}^{l-1})$ instead of $\nu_2(\mathbf{x}^{l-1})$ in Corollary 3, we get that under $\{\mu_2(\mathrm{d}\mathbf{x}^{l-1}) \neq 0\}$:

$$\mathbb{E}_{\theta^l}[\delta\mu_2(\mathrm{d}\mathbf{x}^l)] = 1. \tag{40}$$

Furthermore $\exists B'_l$, independent of $||\boldsymbol{W}^l||_F$, such that $\mathbb{P}_{\theta^l}[B'_l] = 1 - 2^{-N_l}$, and $\exists w'_{\text{min}}, w'_{\text{max}}$ such that under $B'_l \cap \{\mu_2(\mathrm{d}\mathbf{x}^{l-1}) \neq 0\}$:

$$w'_{\text{min}} \sim \frac{1}{N_l R_l}\,\text{Chi-Squared}(1), \quad w'_{\text{max}} \sim \frac{2}{N_l}\,\text{Chi-Squared}(N_l R_l), \qquad w'_{\text{min}} \leq \delta\mu_2(\mathrm{d}\mathbf{x}^l) \leq w'_{\text{max}}. \tag{41}$$

Denoting $A'_l = \bigcap_{k=1}^{l} \Big(B'_k \cap \{\mu_2(\mathrm{d}\mathbf{x}^k) \neq 0\}\Big)$, we also have

$$\mathbb{P}_{\Theta^l}[A'_l] = \prod_{k=1}^{l} \big(1 - 2^{-N_k}\big).$$

Both $\log x$ and $(\log x)^2$ are integrable at 0 since $\int \log x\, dx = x\log x - x$ and $\int (\log x)^2 dx = x(\log x)^2 - 2x\log x + 2x$. By Eq. (39) and Eq. (41), it then follows that $\log\delta\nu_2(\mathbf{x}^l)$ and $\log\delta\mu_2(\mathbf{x}^l)$ have well-defined expectation and variance under $A_l$ and $A'_l$ respectively.

*Now, crucially, let us note that the distributions of $\delta\nu_2(\mathbf{x}^l)$ with respect to $\theta^l|A_l$ and $\delta\mu_2(\mathbf{x}^l)$ with respect to $\theta^l|A'_l$ are fully determined by: (i) the input distributions $P_{\mathbf{x}}(\mathbf{x}) = P_{\mathbf{x}^0}(\mathbf{x}^0)$ and $P_{\mathrm{d}\mathbf{x}}(\mathrm{d}\mathbf{x}) = P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0)$; (ii) the model parameters $\Theta^{l-1}$ up to layer $l-1$.*

We are thus interested in the following infima and suprima:

$$\inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)], \qquad \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)], \tag{42}$$

$$\inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)], \qquad \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)], \tag{43}$$

$$\inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A'_l}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)], \qquad \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A'_l}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)], \tag{44}$$

$$\inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A'_l}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)], \qquad \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A'_l}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)]. \tag{45}$$

Our strategy is to consider:

- Sequences of random variables $(\mathbf{x}^{0,k})_{k\in\mathbb{N}}$, $(\mathrm{d}\mathbf{x}^{0,k})_{k\in\mathbb{N}}$ corresponding to deterministic distributions $P_{\mathbf{x}^{0,k}}(\mathbf{x}^{0,k})$, $P_{\mathrm{d}\mathbf{x}^{0,k}}(\mathrm{d}\mathbf{x}^{0,k})$;
- Sequences of deterministic model parameters $(\Theta^{l-1,k})_{k\in\mathbb{N}}$ up to layer $l-1$;
- Sequences of random variables $(\mathbf{x}^{l-1,k})_{k\in\mathbb{N}}$ and $(\mathrm{d}\mathbf{x}^{l-1,k})_{k\in\mathbb{N}}$ obtained by the simultaneous propagation of $(\mathbf{x}^{0,k}, \mathrm{d}\mathbf{x}^{0,k})$ with parameters $\Theta^{l-1,k}$ up to layer $l-1$;
- Sequences of random variables $(\mathbf{x}^{l,k})_{k\in\mathbb{N}}$ and $(\mathrm{d}\mathbf{x}^{l,k})_{k\in\mathbb{N}}$ obtained by the simultaneous propagation at layer $l$ of $(\mathbf{x}^{l-1,k}, \mathrm{d}\mathbf{x}^{l-1,k})$ with random parameters $(\boldsymbol{\omega}^{l,k}, \boldsymbol{\beta}^{l,k})$;
- Sequences of geometric increments $(\delta\nu_2(\mathbf{x}^{l,k}))_{k\in\mathbb{N}}$ and $(\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k}))_{k\in\mathbb{N}}$, defined as $\delta\nu_2(\mathbf{x}^{l,k}) \equiv \frac{\nu_2(\mathbf{x}^{l,k})}{\nu_2(\mathbf{x}^{l-1,k})}$ and $\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k}) \equiv \frac{\mu_2(\mathrm{d}\mathbf{x}^{l,k})}{\mu_2(\mathrm{d}\mathbf{x}^{l-1,k})}$;
- Sequences of events $(B_{l,k})_{k\in\mathbb{N}}$, $(B'_{l,k})_{k\in\mathbb{N}}$, $(A_{l,k})_{k\in\mathbb{N}}$, $(A'_{l,k})_{k\in\mathbb{N}}$ appropriately defined with respect to $\delta\nu_2(\mathbf{x}^{l,k})$ and $\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k})$.

We will finally consider sequences such that $\mathbb{E}_{\theta^l|A_{l,k}}[-\log\delta\nu_2(\mathbf{x}^{l,k})]$, $\mathrm{Var}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})]$, $\mathbb{E}_{\theta^l|A'_{l,k}}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k})]$, $\mathrm{Var}_{\theta^l|A'_{l,k}}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k})]$ converge to the infima and suprima of Eq. (42), Eq. (43), Eq. (44), Eq. (45) as $k\to\infty$.

*We start by focusing on $\delta\nu_2(\mathbf{x}^l)$ and the reasoning will be easily extended to $\delta\mu_2(\mathrm{d}\mathbf{x}^l)$.*

### D.3.2. Weakly Convergent Subsequence

By Eq. (39), under $B_{l,k} \cap A_{l-1,k}$:

$$\delta\nu_2(\mathbf{x}^{l,k}) \notin ]a,b] \implies \big(a \geq w_{\text{min},k}\big) \vee \big(w_{\text{max},k} > b\big),$$

with $\wedge$ the logical *and*, $\vee$ the logical *or*, and with $w_{\text{min},k}$, $w_{\text{max},k}$ defined as in Eq. (39) with respect to $\delta\nu_2(\mathbf{x}^{l,k})$. Then $\mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin ]a,b]\big] = \mathbb{P}_{\theta^l|B_{l,k}\cap A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin ]a,b]\big]$ can be bounded as

$$\mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin ]a,b]\big] \leq \mathbb{P}_{w\sim\frac{1}{N_l R_l}\text{ Chi-Squared}(1)}\big[w \leq a\big] + \mathbb{P}_{w\sim\frac{2}{N_l}\text{ Chi-Squared}(N_l R_l)}\big[w > b\big].$$

Thus $\forall\epsilon$, $\exists a_\epsilon, b_\epsilon$ such that

$$\begin{aligned} \forall k:\ \mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin ]a_\epsilon,b_\epsilon]\big] &\leq \epsilon, \\ \inf_k \mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \in ]a_\epsilon,b_\epsilon]\big] &\geq 1-\epsilon, \end{aligned}$$

which means that the sequence $\big(\delta\nu_2(\mathbf{x}^{l,k})|A_{l,k}\big)_{k\in\mathbb{N}}$ of random variables is tight. By Lemma 8, it follows that there exists a sequence of strictly increasing indices $(i_k)_{k\in\mathbb{N}}$ and a random variable $X$ such that $\big(\delta\nu_2(\mathbf{x}^{l,i_k})|A_{l,i_k}\big)_{k\in\mathbb{N}}$ converges weakly to $X$: $\delta\nu_2(\mathbf{x}^{l,i_k})|A_{l,i_k} \Rightarrow X$.

If $\mathbb{E}_{\theta^l|A_{l,k}}[-\log\delta\nu_2(\mathbf{x}^{l,k})]$, $\text{Var}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})]$ have well-defined limits equal to the infima and suprima of Eq. (42) and Eq. (43), then $\mathbb{E}_{\theta^l|A_{l,i_k}}[-\log\delta\nu_2(\mathbf{x}^{l,i_k})]$, $\text{Var}_{\theta^l|A_{l,i_k}}[\log\delta\nu_2(\mathbf{x}^{l,i_k})]$ have the same limits. For simplicity of notations and without loss of generality, $(\delta\nu_2(\mathbf{x}^{l,i_k}))_{k\in\mathbb{N}}$ may thus be renamed as $(\delta\nu_2(\mathbf{x}^{l,k}))_{k\in\mathbb{N}}$ such that $\delta\nu_2(\mathbf{x}^{l,k})|A_{l,k} \Rightarrow X$.

We have that for all continuity points $a>0$ of the function $x \mapsto \mathbb{P}[X\leq x]$:

$$\begin{aligned} \mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \leq a\big] &\leq \mathbb{P}_{w\sim\frac{1}{N_l R_l}\text{ Chi-Squared}(1)}\big[w \leq a\big], \\ \mathbb{P}\big[X \leq a\big] &\leq \mathbb{P}_{w\sim\frac{1}{N_l R_l}\text{ Chi-Squared}(1)}\big[w \leq a\big], \end{aligned} \tag{46}$$

where we used the definition of weak convergence: $\mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \leq a\big] \xrightarrow{k\to\infty} \mathbb{P}[X\leq a]$.

Now let us show that the set of discontinuity points of the cumulative distribution function (c.d.f.) $x\mapsto\mathbb{P}[X\leq x]$ on $[0,1]$ has Borel measure equal to 0. Since c.d.f. are always non-decreasing and right-continuous, the set of discontinuity points is the set of non-left-continuity points, i.e. $D=\big\{x\in[0,1]: \lim_{x'\to x^-}\mathbb{P}[X\leq x'] < \mathbb{P}[X\leq x]\big\}$. Let us denote $D_p \equiv \big\{x\in[0,1]: \mathbb{P}[X\leq x] - \lim_{x'\to x^-}\mathbb{P}[X\leq x'] \geq \frac{1}{p}\big\}$. Then the function $\mathbf{1}_{D_p}$ converges point-wise to $\mathbf{1}_D$, i.e. $\forall x\in[0,1]$: $\mathbf{1}_{D_p}(x) \xrightarrow{p\to\infty} \mathbf{1}_D(x)$, and the dominated convergence theorem gives

$$\int_0^1 \mathbf{1}_{D_p}(x)dx \xrightarrow{p\to\infty} \int_0^1 \mathbf{1}_D(x)dx.$$

On the other hand, since $x\mapsto\mathbb{P}[X\leq x]$ is non-decreasing and $0\leq\mathbb{P}[X\leq x]\leq 1$, it follows that $D_p$ is comprised of at most $p$ points, implying that $\int_0^1\mathbf{1}_{D_p}(x)dx=0$. We deduce that $\int_0^1\mathbf{1}_D(x)=0$, i.e. that $D$ has Borel measure equal to 0.

It follows that we can find a sequence of continuity points $a_p>0$ of $x\mapsto\mathbb{P}[X\leq x]$ such that $a_p\xrightarrow{p\to\infty}0$. We then obtain $\mathbb{P}\big[X=0\big]\leq\mathbb{P}\big[X\leq a_p\big]\xrightarrow{p\to\infty}0$ by Eq. (46), and thus $\mathbb{P}\big[X=0\big]=0$. Without loss of generality, we may assume $X>0$ surely (if this is not the case, simply replace $X$ by a constant arbitrary value $>0$ under the zero-probability event $\{X=0\}$).

Now if we consider the function $h$ such that $h(x)=\log x$ if $x>0$, and $h(x)=0$ otherwise, then Lemma 7 implies that $h(\delta\nu_2(\mathbf{x}^{l,k}))|A_{l,k}\Rightarrow h(X)$, i.e. $\log\delta\nu_2(\mathbf{x}^{l,k})|A_{l,k}\Rightarrow\log X$. If we consider $h(x)=x^2$, we further deduce that $\delta\nu_2(\mathbf{x}^{l,k})^2|A_{l,k}\Rightarrow X^2$ and that $\big(\log\delta\nu_2(\mathbf{x}^{l,k})\big)^2|A_{l,k}\Rightarrow(\log X)^2$.

### D.3.3. UNIFORM INTEGRABILITY

Since both $x \mapsto \mathbf{1}_{\{x \geq M\}} x$ and $x \mapsto \mathbf{1}_{\{x^2 \geq M\}} x^2$ are non-decreasing for $x > 0$, Eq. (39) implies that

$$\sup_k \mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{\delta\nu_2(\mathbf{x}^{l,k}) \geq M\}} \delta\nu_2(\mathbf{x}^{l,k}) \Big] \leq \mathbb{E}_{w \sim \frac{2}{N_l} \text{ Chi-Squared}(N_l R_l)} \Big[ \mathbf{1}_{\{w \geq M\}} w \Big] \xrightarrow{M \to \infty} 0,$$
$$\sup_k \mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{\delta\nu_2(\mathbf{x}^{l,k})^2 \geq M\}} \delta\nu_2(\mathbf{x}^{l,k})^2 \Big] \leq \mathbb{E}_{w \sim \frac{2}{N_l} \text{ Chi-Squared}(N_l R_l)} \Big[ \mathbf{1}_{\{w^2 \geq M\}} w^2 \Big] \xrightarrow{M \to \infty} 0.$$

Since $\delta\nu_2(\mathbf{x}^{l,k}) \geq 0$, it follows that both $\big(\delta\nu_2(\mathbf{x}^{l,k}) | A_{l,k}\big)_{k \in \mathbb{N}}$ and $\big(\delta\nu_2(\mathbf{x}^{l,k})^2 | A_{l,k}\big)_{k \in \mathbb{N}}$ are uniformly integrable, implying by Lemma 9 that

$$\mathbb{E}_{\theta^l | A_{l,k}} \big[ \delta\nu_2(\mathbf{x}^{l,k}) \big] \xrightarrow{k \to \infty} \mathbb{E}\big[X\big], \qquad \mathbb{E}_{\theta^l | A_{l,k}} \big[ \delta\nu_2(\mathbf{x}^{l,k})^2 \big] \xrightarrow{k \to \infty} \mathbb{E}\big[X^2\big].$$

Again since $x \mapsto \mathbf{1}_{\{x \geq M\}} x$ is non-decreasing for $x > 0$, Eq. (39) implies that under $B_{l,k} \cap \big\{w_{\text{min},k} > 0\big\} \cap \big\{w_{\text{max},k} > 0\big\}$:

$$\log w_{\text{min},k} \leq \log \delta\nu_2(\mathbf{x}^{l,k}) \leq \log w_{\text{max},k},$$
$$\big| \log \delta\nu_2(\mathbf{x}^{l,k}) \big| \leq \max \Big( \big| \log w_{\text{min},k} \big|, \big| \log w_{\text{max},k} \big| \Big),$$
$$\mathbf{1}_{\{| \log \delta\nu_2(\mathbf{x}^{l,k}) | \geq M\}} \big| \log \delta\nu_2(\mathbf{x}^{l,k}) \big| \leq \max \Big( \mathbf{1}_{\{| \log w_{\text{min},k} | \geq M\}} \big| \log w_{\text{min},k} \big|, \mathbf{1}_{\{| \log w_{\text{max},k} | \geq M\}} \big| \log w_{\text{max},k} \big| \Big),$$
$$\mathbf{1}_{\{| \log \delta\nu_2(\mathbf{x}^{l,k}) | \geq M\}} \big| \log \delta\nu_2(\mathbf{x}^{l,k}) \big| \leq \mathbf{1}_{\{| \log w_{\text{min},k} | \geq M\}} \big| \log w_{\text{min},k} \big| + \mathbf{1}_{\{| \log w_{\text{max},k} | \geq M\}} \big| \log w_{\text{max},k} \big|.$$

Similarly, we have that under $B_{l,k} \cap \big\{w_{\text{min},k} > 0\big\} \cap \big\{w_{\text{max},k} > 0\big\}$:

$$\mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \big( \log \delta\nu_2(\mathbf{x}^{l,k}) \big)^2 \leq \max \Big( \mathbf{1}_{\{(\log w_{\text{min},k})^2 \geq M\}} \big( \log w_{\text{min},k} \big)^2, \mathbf{1}_{\{(\log w_{\text{max},k})^2 \geq M\}} \big( \log w_{\text{max},k} \big)^2 \Big),$$
$$\mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \big( \log \delta\nu_2(\mathbf{x}^{l,k}) \big)^2 \leq \mathbf{1}_{\{(\log w_{\text{min},k})^2 \geq M\}} \big( \log w_{\text{min},k} \big)^2 + \mathbf{1}_{\{(\log w_{\text{max},k})^2 \geq M\}} \big( \log w_{\text{max},k} \big)^2.$$

Using $\mathbb{P}_{\theta^l}[w_{\text{min},k} = 0] = 0$ and $\mathbb{P}_{\theta^l}[w_{\text{max},k} = 0] = 0$, and denoting $\text{Chi-Squared}(1)^*$ and $\text{Chi-Squared}(N_l R_l)^*$ the chi-squared distributions excluding zero values, we get that

$$\mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{| \log \delta\nu_2(\mathbf{x}^{l,k}) | \geq M\}} \big| \log \delta\nu_2(\mathbf{x}^{l,k}) \big| \Big]$$
$$\leq \mathbb{E}_{w \sim \frac{1}{N_l R_l} \text{ Chi-Squared}(1)^*} \Big[ \mathbf{1}_{\{| \log w | \geq M\}} \big| \log w \big| \Big] + \mathbb{E}_{w \sim \frac{2}{N_l} \text{ Chi-Squared}(N_l R_l)^*} \Big[ \mathbf{1}_{\{| \log w | \geq M\}} \big| \log w \big| \Big],$$
$$\mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \big( \log \delta\nu_2(\mathbf{x}^{l,k}) \big)^2 \Big]$$
$$\leq \mathbb{E}_{w \sim \frac{1}{N_l R_l} \text{ Chi-Squared}(1)^*} \Big[ \mathbf{1}_{\{(\log w)^2 \geq M\}} \big( \log w \big)^2 \Big] + \mathbb{E}_{w \sim \frac{2}{N_l} \text{ Chi-Squared}(N_l R_l)^*} \Big[ \mathbf{1}_{\{(\log w)^2 \geq M\}} \big( \log w \big)^2 \Big].$$

It follows that

$$\sup_k \mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{| \log \delta\nu_2(\mathbf{x}^{l,k}) | \geq M\}} \big| \log \delta\nu_2(\mathbf{x}^{l,k}) \big| \Big] \xrightarrow{M \to \infty} 0,$$
$$\sup_k \mathbb{E}_{\theta^l | A_{l,k}} \Big[ \mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \big( \log \delta\nu_2(\mathbf{x}^{l,k}) \big)^2 \Big] \xrightarrow{M \to \infty} 0.$$

Thus both $\big( \log \delta\nu_2(\mathbf{x}^{l,k}) | A_{l,k} \big)_{k \in \mathbb{N}}$ and $\big( (\log \delta\nu_2(\mathbf{x}^{l,k}))^2 | A_{l,k} \big)_{k \in \mathbb{N}}$ are uniformly integrable, and by Lemma 9:

$$\mathbb{E}_{\theta^l | A_{l,k}} \big[ \log \delta\nu_2(\mathbf{x}^{l,k}) \big] \xrightarrow{k \to \infty} \mathbb{E}\big[ \log X \big], \qquad \mathbb{E}_{\theta^l | A_{l,k}} \big[ \big( \log \delta\nu_2(\mathbf{x}^{l,k}) \big)^2 \big] \xrightarrow{k \to \infty} \mathbb{E}\big[ (\log X)^2 \big].$$

### D.3.4. Bounding Moments of $\delta\nu_2(\mathbf{x}^{l,k})$

First let us bound $\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]$ from above. For each channel, the variance is bounded as

$$\begin{aligned}\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})}{\nu_2(\mathbf{x}^{l-1,k})}\right] &\leq \mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})^2}{\nu_2(\mathbf{x}^{l-1,k})^2}\right] \leq \mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{y}^{l,k})^2}{\nu_2(\mathbf{x}^{l-1,k})^2}\right]\\ &\leq R_l^2\,\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\sum\nolimits_{i\neq i'}\big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2\big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i'}\big)^2\hat{\lambda}_i\hat{\lambda}_{i'}+\sum\nolimits_i\big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^4\hat{\lambda}_i^2\Big]\\ &\leq R_l^2\sum\nolimits_{i,i'}\left(\frac{2}{R_l}\right)\left(\frac{2}{R_l}\right)3\hat{\lambda}_i\hat{\lambda}_{i'}=12.\end{aligned}$$

Since the different channels are independent, we get that

$$\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]=\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[\frac{1}{N_l}\sum_{\mathrm{c}}\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})}{\nu_2(\mathbf{x}^{l-1,k})}\right]=\frac{1}{N_l^2}\sum_{\mathrm{c}}\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})}{\nu_2(\mathbf{x}^{l-1,k})}\right]\leq\frac{12}{N_l}.$$

Next we bound $\big|\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]-1\big|$. Using $\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]=1$ by Eq. (38):

$$\begin{aligned}&\big|\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]-1\big|\\ &\quad=\big|\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]-\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]\big|\\ &\quad=\left|\left(\frac{1}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}-1\right)\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\mathbf{1}_{A_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\big]-\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\mathbf{1}_{A^{\mathrm{c}}_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\big]\right|\\ &\quad\leq\frac{\mathbb{P}_{\theta^l|A_{l-1,k}}[A^{\mathrm{c}}_{l,k}]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}\big|\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\mathbf{1}_{A_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\big]\big|+\mathbb{P}_{\theta^l|A_{l-1,k}}\big[A^{\mathrm{c}}_{l,k}\big]^{\frac{1}{2}}\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})^2\big]^{\frac{1}{2}} &&(47)\\ &\quad\leq\left(\frac{\mathbb{P}_{\theta^l|A_{l-1,k}}[A^{\mathrm{c}}_{l,k}]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}+\mathbb{P}_{\theta^l|A_{l-1,k}}\big[A^{\mathrm{c}}_{l,k}\big]^{\frac{1}{2}}\right)\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})^2\big]^{\frac{1}{2}} &&(48)\\ &\quad\leq\left(\frac{2^{-N_l}}{1-2^{-N_l}}+2^{-\frac{N_l}{2}}\right)\Big(1+\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]\Big)^{\frac{1}{2}}\leq\epsilon_l\left(1+\frac{12}{N_l}\right)^{\frac{1}{2}}\leq 2\epsilon_l, &&(49)\end{aligned}$$

where we applied Cauchy-Schwarz inequality in Eq. (47) and Eq. (48), defined $\epsilon_l\equiv\frac{2^{-N_l}}{1-2^{-N_l}}+2^{-\frac{N_l}{2}}$ and used $\Big(1+\frac{12}{N_l}\Big)^{\frac{1}{2}}\leq 2$ under the large width assumption.

We are then able to bound $\mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]$ from above:

$$\begin{aligned}\mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]&=\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})^2\big]-\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]^2\\ &\leq\frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\mathbf{1}_{A_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})^2\big]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}-1+1-\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]^2\\ &\leq\frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})^2\big]-1}{1-2^{-N_l}}+\frac{1}{1-2^{-N_l}}-1+\big|\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]-1\big|\big|\mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]+1\big|\\ &\leq\frac{\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]}{1-2^{-N_l}}+\frac{2^{-N_l}}{1-2^{-N_l}}+2\epsilon_l\left(\frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]}{1-2^{-N_l}}+1\right)\\ &\leq\frac{1}{1-2^{-N_l}}\left(\frac{12}{N_l}+2^{-N_l}\right)+2\epsilon_l\left(\frac{1}{1-2^{-N_l}}+1\right)\leq\frac{24}{N_l}, &&(50)\end{aligned}$$

where we used again the fact that the terms in $2^{-N_l}$ are negligible with respect to $\frac{12}{N_l}$ under the large width assumption.

Finally let us bound $\mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]$ from below. In the remaining of this calculation, the conditionality on $\big\{||\boldsymbol{W}^l||_F^2>0\big\}$ is assumed but omitted for simplicity of notation. This conditionality has no effect on expectations and probabilities since $\big\{||\boldsymbol{W}^l||_F^2>0\big\}$ has probability one.

We first note that $\Big(\mathbf{1}_{B_{l,k}},\, \frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\Big)$ is fully determined by $\Big((\tilde{v}^l_\mathrm{c})_{1\leq\mathrm{c}\leq N_l},\, \frac{\boldsymbol{W}^l}{||\boldsymbol{W}^l||_F^2}\Big)$, which is itself independent from $||\boldsymbol{W}^l||_F^2$. It follows that $\Big(\mathbf{1}_{B_{l,k}},\, \frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\Big)$ is independent from $||\boldsymbol{W}^l||_F^2$, and thus that

$$\begin{aligned}
&\mathrm{Var}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]\\
&\quad= \frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\mathbf{1}_{B_{l,k}}\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}||\boldsymbol{W}^l||_F^4\Big]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[B_{l,k}]} - \frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\mathbf{1}_{B_{l,k}}\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}||\boldsymbol{W}^l||_F^2\Big]^2}{\mathbb{P}_{\theta^l|A_{l-1,k}}[B_{l,k}]^2}\\
&\quad= \frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\mathbf{1}_{B_{l,k}}\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\Big]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[B_{l,k}]}\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^4\big] - \frac{\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\mathbf{1}_{B_{l,k}}\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\Big]^2}{\mathbb{P}_{\theta^l|A_{l-1,k}}[B_{l,k}]^2}\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]^2\\
&\quad= \mathbb{E}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\right]\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^4\big] - \mathbb{E}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]^2\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]^2\\
&\quad\geq \mathbb{E}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]^2\Big(\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^4\big] - \mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]^2\Big)\\
&\quad\geq \mathbb{E}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]^2\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big].
\end{aligned}$$

Due to $\mathbb{P}_{\theta^l|A_{l-1,k}\cap B_{l,k}}[A_{l,k}]=1$ and $A_{l,k}\subseteq A_{l-1,k}\cap B_{l,k}$, the conditionality on $A_{l-1,k}\cap B_{l,k}$ can be replaced by the conditionality on $A_{l,k}$:

$$\mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big] \geq \mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]^2\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]. \tag{51}$$

It remains to bound the terms $\mathbb{E}_{\theta^l|A_{l,k}}\Big[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\Big]^2$ and $\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]$. A computation similar to Eq. (48) gives

$$\left|\mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right] - \mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]\right| \leq \epsilon_l\,\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\right]^{\frac{1}{2}}. \tag{52}$$

The term $\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\Big]^{\frac{1}{2}}$ of Eq. (52) can be bounded using Eq. (37):

$$\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2} = \frac{R_l}{N_l}\sum_\mathrm{c}u^l_\mathrm{c}\sum_i\big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2\hat{\lambda}_i\frac{1}{||\boldsymbol{W}^l||_F^2} \leq \frac{R_l}{N_l}\sum_\mathrm{c}\sum_i\big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2\frac{1}{||\boldsymbol{W}^l||_F^2} = \frac{R_l}{N_l},$$

$$\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\right]^{\frac{1}{2}} \leq \frac{R_l}{N_l}.$$

As for the term $\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\Big]$ of Eq. (52), we get by independence of $||\boldsymbol{W}^l||_F$ and $\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}$ that

$$\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big] = \mathbb{E}_{\theta^l|A_{l-1,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big] = 1,$$

$$\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right] = \frac{1}{\mathbb{E}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]} = \frac{1}{\frac{2}{R_l}N_lR_l} = \frac{1}{2N_l}.$$

We have $\epsilon_l = \frac{2^{-N_l}}{1-2^{-N_l}} + 2^{-\frac{N_l}{2}} \ll \frac{1}{2N_{l-1}n^d} \leq \frac{1}{2R_l}$ under the large width assumption. Then, by Eq. (52):

$$\left| \mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right] - \frac{1}{2N_l} \right| \ll \frac{1}{2R_l}\frac{R_l}{N_l} = \frac{1}{2N_l}, \tag{53}$$

$$\mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right] \geq \frac{1}{4N_l}. \tag{54}$$

The variance $\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big]$ is given by

$$\begin{aligned}
\mathrm{Var}_{\theta^l|A_{l-1,k}}\big[||\boldsymbol{W}^l||_F^2\big] &= \left(\frac{2}{R_l}\right)^2 \left( \mathbb{E}_{\theta^l|A_{l-1,k}}\left[ \sum\nolimits_{(\mathrm{c},i),(\mathrm{c}',i')} \left(\frac{2}{R_l}\right)^{-1} \big(\boldsymbol{W}^l_{\mathrm{c},i}\big)^2 \left(\frac{2}{R_l}\right)^{-1} \big(\boldsymbol{W}^l_{\mathrm{c}',i'}\big)^2 \right] - N_l^2 R_l^2 \right) \\
&= \left(\frac{2}{R_l}\right)^2 \left( \Big( \sum\nolimits_{(\mathrm{c},i)\neq(\mathrm{c}',i')} 1 \Big) + \Big( \sum\nolimits_{(\mathrm{c},i)} 3 \Big) - N_l^2 R_l^2 \right) = \frac{8N_l}{R_l}.
\end{aligned} \tag{55}$$

Finally combining Eq. (51), Eq. (54) and Eq. (55):

$$\mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big] \geq \left(\frac{1}{4N_l}\right)^2 \frac{8N_l}{R_l} = \frac{1}{2N_l R_l}. \tag{56}$$

### D.3.5. Consequence for $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$

Using Eq. (49) and taking the limit $k \to \infty$:

$$\begin{aligned}
\Big| \mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big] - 1 \Big| &\leq 2\epsilon_l, \\
\Big| \mathbb{E}\big[X\big] - 1 \Big| &\leq 2\epsilon_l.
\end{aligned}$$

Similarly, using Eq. (50) and Eq. (56) and taking the limit $k \to \infty$:

$$\begin{aligned}
\frac{1}{2N_l R_l} \leq \mathrm{Var}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big] = \mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})^2\big] - \mathbb{E}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k})\big]^2 &\leq \frac{24}{N_l}, \\
\frac{1}{2N_l R_l} \leq \mathrm{Var}\big[X\big] = \mathbb{E}\big[X^2\big] - \mathbb{E}\big[X\big]^2 &\leq \frac{24}{N_l}.
\end{aligned}$$

Thus $\big|\mathbb{E}\big[X\big]-1\big|$ is exponentially small in $N_l$, while the standard deviation of $X$ behaves as a power-law of $N_l$: $\frac{1}{\sqrt{2N_l R_l}} \leq \mathrm{Var}\big[X\big]^{\frac{1}{2}} \leq \sqrt{\frac{24}{N_l}}$. This means that $\big|\mathbb{E}\big[X\big]-1\big|$ is much smaller than the effect of the log-concavity:

$$\begin{aligned}
\big|\mathbb{E}\big[X\big]-1\big| \ll \log \mathbb{E}\big[X\big] - \mathbb{E}\big[\log X\big] \leq \mathbb{E}\big[X\big] - 1 - \mathbb{E}\big[\log X\big] \quad &\Longrightarrow \quad \big|\mathbb{E}\big[X\big]-1\big| < \mathbb{E}\big[X\big] - 1 - \mathbb{E}\big[\log X\big] \\
&\Longrightarrow \quad \big|\mathbb{E}\big[X\big]-1\big| - \big(\mathbb{E}\big[X\big]-1\big) < -\mathbb{E}\big[\log X\big] \\
&\Longrightarrow \quad 0 < \mathbb{E}\big[-\log X\big].
\end{aligned}$$

In addition, $X$ has small standard deviation around $\mathbb{E}\big[X\big]$ since $\mathrm{Var}\big[X\big]^{\frac{1}{2}} \ll 1$ under the large width assumption. This implies that

$$\begin{aligned}
0 < \lim_{k\to\infty} \mathbb{E}_{\theta^l|A_{l,k}}[-\log \delta\nu_2(\mathbf{x}^{l,k})] &= \mathbb{E}\big[-\log X\big] \ll 1, \\
0 < \lim_{k\to\infty} \mathrm{Var}_{\theta^l|A_{l,k}}[\log \delta\nu_2(\mathbf{x}^{l,k})] &= \mathrm{Var}\big[\log X\big] \ll 1.
\end{aligned}$$

Now if we alternately consider sequences $\big(\delta\nu_2(\mathbf{x}^{l,k})|A_{l,k}\big)_{k\in\mathbb{N}}$ corresponding to distributions $P_{\mathbf{x}^{0,k}}(\mathbf{x}^{0,k})$, $P_{\mathrm{d}\mathbf{x}^{0,k}}(\mathrm{d}\mathbf{x}^{0,k})$, and parameters $\Theta^{l-1,k}$ up to layer $l-1$, such that

$$\begin{aligned}
\lim_{k\to\infty} \mathbb{E}_{\theta^l|A_{l,k}}[-\log\delta\nu_2(\mathbf{x}^{l,k})] &= \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)],\\
\lim_{k\to\infty} \mathbb{E}_{\theta^l|A_{l,k}}[-\log\delta\nu_2(\mathbf{x}^{l,k})] &= \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)],\\
\lim_{k\to\infty} \mathrm{Var}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})] &= \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)],\\
\lim_{k\to\infty} \mathrm{Var}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})] &= \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)],
\end{aligned}$$

then we obtain that

$$\begin{aligned}
0 &< \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)] \ll 1.
\end{aligned}$$

The final remaining dependency is the dependency in $N_l$ and $R_l$. Since $R_l = K_l^d N_{l-1} \leq n^d N_{l-1}$, and since $(N_l)_{l\in\mathbb{N}}$ is bounded, it follows that $(R_l)_{l\in\mathbb{N}}$ is also bounded. If we denote

$$\begin{gathered}
N_{\text{min}} \equiv \min_l N_l, \quad N_{\text{max}} \equiv \max_l N_l, \quad R_{\text{min}} \equiv \min_l R_l, \quad R_{\text{max}} \equiv \max_l R_l,\\
\mathcal{I}_N \equiv \{N_{\text{min}},\dots,N_{\text{max}}\}, \quad \mathcal{I}_R \equiv \{R_{\text{min}},\dots,R_{\text{max}}\},
\end{gathered}$$

then we finally get

$$\begin{aligned}
0 &< \min_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \max_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \min_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)] \ll 1,\\
0 &< \max_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)] \ll 1.
\end{aligned}$$

The whole reasoning can immediately be transposed to $\mu_2(\mathrm{d}\mathbf{x}^l)$ to get

$$\begin{aligned}
0 &< \min_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A'_l}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \ll 1,\\
0 &< \max_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathbb{E}_{\theta^l|A'_l}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \ll 1,\\
0 &< \min_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \inf_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A'_l}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \ll 1,\\
0 &< \max_{N_l\in\mathcal{I}_N,R_l\in\mathcal{I}_R} \sup_{P_{\mathbf{x}^0}(\mathbf{x}^0),P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0),\Theta^{l-1}} \mathrm{Var}_{\theta^l|A'_l}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \ll 1.
\end{aligned}$$

It follows that there exists small positive constants $1 \gg m_{\text{min}}, m_{\text{max}}, v_{\text{min}}, v_{\text{max}} > 0$ such that $\forall l$:

$$m_{\text{min}} \leq \mathbb{E}_{\theta^l|A_l}[-\log\delta\nu_2(\mathbf{x}^l)] \quad \leq m_{\text{max}}, \qquad v_{\text{min}} \leq \mathrm{Var}_{\theta^l|A_l}[\log\delta\nu_2(\mathbf{x}^l)] \quad \leq v_{\text{max}}, \tag{57}$$

$$m_{\text{min}} \leq \mathbb{E}_{\theta^l|A'_l}[-\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \leq m_{\text{max}}, \qquad v_{\text{min}} \leq \mathrm{Var}_{\theta^l|A'_l}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^l)] \leq v_{\text{max}}. \tag{58}$$

### D.3.6. PROOF CONCLUSION

Again we start by focusing on $\delta\nu_2(\mathbf{x}^l)$ and the reasoning will easily be extended to $\delta\mu_2(\mathrm{d}\mathbf{x}^l)$. Let us define under $A_k$:

$$X_k \equiv \log \delta\nu_2(\mathbf{x}^k), \qquad Y_k \equiv \mathbb{E}_{\theta^k|A_k}[\log \delta\nu_2(\mathbf{x}^k)], \qquad Z_k \equiv \log \delta\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k|A_k}[\log \delta\nu_2(\mathbf{x}^k)].$$

Using Eq. (57), we have that under $A_k$:

$$\begin{aligned} m_{\text{min}} \leq -\, Y_k \leq m_{\text{max}}, &\qquad v_{\text{min}} \leq \mathrm{Var}_{\theta^k|A_k}[Z_k] \leq v_{\text{max}}, \\ -m_{\text{max}} \leq Y_k \leq -m_{\text{min}}, &\qquad v_{\text{min}} \leq \mathrm{Var}_{\theta^k|A_k}[Z_k] \leq v_{\text{max}}. \end{aligned}$$

By Lemma 10, we deduce that there exist random variables $m_l$, $s_l$ such that under $A_l$:

$$\begin{aligned} \sum_{k=1}^{l} \log \delta\nu_2(\mathbf{x}^k) &= l m_l + \sqrt{l} s_l, \quad -m_{\text{max}} \leq m_l \leq -m_{\text{min}}, \quad \mathbb{E}_{\Theta^l|A_l}[s_l] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\text{max}}, \\ \log\left(\frac{\nu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^0)}\right) &= l m_l + \sqrt{l} s_l, \quad -m_{\text{max}} \leq m_l \leq -m_{\text{min}}, \quad \mathbb{E}_{\Theta^l|A_l}[s_l] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\text{max}}. \end{aligned}$$

Finally changing the variable $m_l$ to $-m_l$, we get that under $A_l$:

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

Applying the exact same reasoning to $\mu_2(\mathrm{d}\mathbf{x}^l)$, we deduce that there exist random variables $m'_l$, $s'_l$ such that under $A'_l$:

$$\log\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\right) = -l m'_l + \sqrt{l} s'_l, \quad m_{\text{min}} \leq m'_l \leq m_{\text{max}}, \quad \mathbb{E}_{\Theta^l|A'_l}[s'_l] = 0, \quad v_{\text{min}} \leq \mathrm{Var}_{\Theta^l|A'_l}[s'_l] \leq v_{\text{max}}.$$

### D.3.7. ILLUSTRATION

Let us give an illustration in the fully-connected case with constant width, $N_l = N = 100$ and $R_l = N = 100$. The bounds $m_{\text{min}}$, $m_{\text{max}}$, $v_{\text{min}}$, $v_{\text{max}}$ are obtained by considering the extreme cases for $u^l_{\text{c}}$ and $R_l \sum_i (\hat{\boldsymbol{W}}^l_{\text{c},i})^2 \hat{\lambda}_i$ in Eq. (37):

- We obtain *minimum bounds* by considering $u^l_{\text{c}} \sim 1/2$ and $R_l \sum_i (\hat{\boldsymbol{W}}^l_{\text{c},i})^2 \hat{\lambda}_i \sim 2\,\text{Chi-Squared}(N)/N$, leading to $\delta\nu_2(\mathbf{x}^l)$, $\delta\mu_2(\mathrm{d}\mathbf{x}^l) \sim \text{Chi-Squared}(N^2)/N^2$;
- We obtain *maximum bounds* by considering $u^l_{\text{c}} \sim \text{Bernouilli}(1/2)$ and $R_l \sum_i (\hat{\boldsymbol{W}}^l_{\text{c},i})^2 \hat{\lambda}_i \sim 2\,\text{Chi-Squared}(1)$.

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.

## D.4. The Conditionality on $A_l$ is Highly Negligible

The events $A_l$, $A'_l$ defined in Theorem 1 have probabilities equal to $\prod_{k=1}^{l} \left(1 - 2^{-N_k}\right)$. Thus

$$-\mathbb{P}_{\Theta^l}[A_l^{\text{c}}] \simeq \log\left(1 - \mathbb{P}_{\Theta^l}[A_l^{\text{c}}]\right) = \log \mathbb{P}_{\Theta^l}[A_l] = \sum\nolimits_{k=1}^{l} \log\left(1 - 2^{-N_k}\right) \simeq -\sum\nolimits_{k=1}^{l} 2^{-N_k},$$

implying that $\mathbb{P}_{\Theta^l}[A_l^{\text{c}}] = \mathbb{P}_{\Theta^l}[A_l'^{\text{c}}] \simeq \sum_{k=1}^{l} 2^{-N_k}$. It follows that $\mathbb{P}_{\Theta^l}[A_l^{\text{c}}]$, $\mathbb{P}_{\Theta^l}[A_l'^{\text{c}}]$ grow linearly in the depth but decay exponentially in the width.

In practice, $\mathbb{P}_{\Theta^l}[A_l^{\text{c}}]$, $\mathbb{P}_{\Theta^l}[A_l'^{\text{c}}]$ are thus highly negligible and the conditionality on $A_l$, $A'_l$ is also highly negligible. For example, in the case of constant width $N_l = 100$ and total depth $L = 200$, we numerically find $\mathbb{P}_{\Theta^L}[A_L^{\text{c}}] = \mathbb{P}_{\Theta^L}[A_L'^{\text{c}}] \simeq 3.2 \times 10^{-28}$.

### D.5. Relation to the Terms $\overline{m}$, $\underline{m}$, $\underline{s}$ of Section 4

Here we relate Theorem 1 to the terms $\overline{m}$, $\underline{m}$, $\underline{s}$ defined in Section 4, under the conditionality $A_k$, $A'_k$. By Eq. (49), we have that $|\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \leq 2\epsilon_k \ll 1$. This implies that under $A_k$:

$$|\overline{m}[\nu_2(\mathbf{x}^k)]| = \big|\log \mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)]\big| \simeq \big|\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] - 1\big| \leq 2\epsilon_k.$$

Similarly, we have that $|\mathbb{E}_{\theta^k|A'_k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)] - 1| \leq 2\epsilon_k \ll 1$, and that under $A'_k$:

$$|\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^k)]| = \big|\log \mathbb{E}_{\theta^k|A'_k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)]\big| \simeq \big|\mathbb{E}_{\theta^k|A'_k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)] - 1\big| \leq 2\epsilon_k.$$

The terms $\overline{m}[\nu_2(\mathbf{x}^k)]$, $\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^k)]$ are thus exponentially small in $N_k$, implying that the evolution with depth of $\nu_2(\mathbf{x}^l), \mu_2(\mathrm{d}\mathbf{x}^l)$ is dominated by the negative drift terms: $\underline{m}[\nu_2(\mathbf{x}^l)] < 0, \underline{m}[\mu_2(\mathrm{d}\mathbf{x}^l)] < 0$ and the diffusion terms: $\underline{s}[\nu_2(\mathbf{x}^l)]$, $\underline{s}[\mu_2(\mathrm{d}\mathbf{x}^l)]$.

### D.6. Proof of Theorem 2

**Theorem 2** (normalized sensitivity increments of vanilla nets)**.** *Denoting* $\mathbf{y}^{l,\pm} \equiv \max\big(\pm\mathbf{y}^l, 0\big)$*, the dominating term under* $\{\mu_2(\mathbf{x}^{l-1}) > 0\}$ *in the evolution of* $\chi^l$ *is*

$$\delta\chi^l \simeq \exp(\overline{m}[\chi^l]) = \underbrace{\left(1 - \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,-}\big)}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-\frac{1}{2}}}_{\in[1,\sqrt{2}]}.$$

**Proof.** The dominating term in the evolution of $\chi^l$ is given by

$$\delta\chi^l \simeq \exp(\overline{m}[\chi^l]) = \left(\frac{\mathbb{E}_{\theta^l}[\delta\mu_2(\mathrm{d}\mathbf{x}^l)]}{\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{x}^l)]}\right)^{\frac{1}{2}}. \tag{59}$$

First we consider the term $\mathbb{E}_{\theta^l}[\delta\mu_2(\mathbf{x}^l)]$. Again we use the definitions and notations from Section B. We further denote $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \dots, \lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^l \equiv \boldsymbol{W}^l(\boldsymbol{e}_1, \dots, \boldsymbol{e}_{R_l})$. Using these notations, we get that $\forall$c:

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{y}^l) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\hat{\varphi}(\mathbf{y}^l, \boldsymbol{\alpha})^2_\mathrm{c}\Big] &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^l_{\mathrm{c},:}\hat{\rho}(\mathbf{x}^{l-1}, \boldsymbol{\alpha})\big)^2\Big] \\ &= \sum\nolimits_i \big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2 \lambda_i.\end{aligned} \tag{60}$$

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

$$\mathbb{E}_{\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)] = \frac{2}{R_l}\sum\nolimits_i \lambda_i = \frac{2}{R_l}\operatorname{Tr}\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})] = 2\mu_2(\mathbf{x}^{l-1}). \tag{61}$$

where we used Corollary 3 in Eq. (61). The symmetric propagation gives

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[(\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2\big] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{y}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}\big]^2 + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[(\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2\big] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{y}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}\big]^2 \\ &= \nu_{2,\mathrm{c}}(\mathbf{y}^{l,+}) - \nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})^2 + \nu_{2,\mathrm{c}}(\mathbf{y}^{l,-}) - \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})^2 \\ &= \nu_{2,\mathrm{c}}(\mathbf{y}^l) - \Big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})^2 + \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})^2\Big).\end{aligned} \tag{62}$$

Since $\mathbf{y}^l = \mathbf{y}^{l,+} - \mathbf{y}^{l,-}$ and $|\mathbf{y}^l| = \mathbf{y}^{l,+} + \mathbf{y}^{l,-}$, we can express $\nu_{1,\mathrm{c}}(\mathbf{y}^l)$ and $\nu_{1,\mathrm{c}}(|\mathbf{y}^l|)$ as

$$\nu_{1,\mathrm{c}}(\mathbf{y}^l)^2 = \big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}) - \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big)^2 = \nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})^2 + \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})^2 - 2\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}), \tag{63}$$

$$\nu_{1,\mathrm{c}}(|\mathbf{y}^l|)^2 = \big(\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+}) + \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})\big)^2 = \nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})^2 + \nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})^2 + 2\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}). \tag{64}$$

Using Eq. (63), we can then rewrite Eq. (62) as

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) &= \nu_{2,\mathrm{c}}(\mathbf{y}^l) - \nu_{1,\mathrm{c}}(\mathbf{y}^l)^2 - 2\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) \\ &= \mu_{2,\mathrm{c}}(\mathbf{y}^l) - 2\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}).\end{aligned} \tag{65}$$

Combining Eq. (61) and Eq. (65):

$$\begin{aligned}\mathbb{E}_{\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})], \\ 2\mathbb{E}_{\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= 2\mu_2(\mathbf{x}^{l-1}) - 2\mathbb{E}_{\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})], \\ \mathbb{E}_{\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{x}^l)] &= \mu_2(\mathbf{x}^{l-1})\left(1 - \mathbb{E}_{\theta^l}\left[\frac{\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right]\right).\end{aligned} \tag{66}$$

where Eq. (66) was obtained by symmetry of the propagation. We then get

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

Combining with Eq. (59) and $\mathbb{E}_{\theta^l}[\delta\mu_2(\mathrm{d}\mathbf{x}^l)] = 1$ by Eq. (40) in the proof of Theorem 1, we finally get

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

To obtain the bounds on $\exp(\overline{m}[\chi^l])$, we use Eq. (63) and Eq. (64):

$$\begin{aligned}4\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) + \nu_{1,\mathrm{c}}(\mathbf{y}^l)^2 = \nu_{1,\mathrm{c}}(|\mathbf{y}^l|)^2 &\leq \nu_{2,\mathrm{c}}(|\mathbf{y}^l|) = \nu_{2,\mathrm{c}}(\mathbf{y}^l), \\ 4\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) &\leq \nu_{2,\mathrm{c}}(\mathbf{y}^l) - \nu_{1,\mathrm{c}}(\mathbf{y}^l)^2 = \mu_{2,\mathrm{c}}(\mathbf{y}^l).\end{aligned} \tag{67}$$

Given $\mathbb{E}_{\theta^l}[\mu_{2,\mathrm{c}}(\mathbf{y}^l)] = 2\mu_2(\mathbf{x}^{l-1})$ by Eq. (61), we deduce that $4\mathbb{E}_{\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})] \leq 2\mu_2(\mathbf{x}^{l-1})$, and thus that

$$\begin{aligned}\mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,-}\big)}{\mu_2(\mathbf{x}^{l-1})}\right] &\leq \frac{1}{2}, \\ 1 \leq \exp(\overline{m}[\chi^l]) = \left(1 - \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{y}^{l,-}\big)}{\mu_2(\mathbf{x}^{l-1})}\right]\right)^{-\frac{1}{2}} &\leq \sqrt{2}.\end{aligned}$$

□

### D.7. If the Drift of $\chi^l$ Is Larger than Diffusion and if $\nu_2(\mathbf{x}^l)$, $\mu_2(\mathrm{d}\mathbf{x}^l)$ are Lognormal, then $\mu_2(\mathbf{x}^l)\,/\,\nu_2(\mathbf{x}^l)\to 0$ a.s.

**Lemma 11.** *For a sequence of random variables $(X_l)_{l\in\mathbb{N}}$ 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 $N_\epsilon$ 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 that $\mathbb{E}[N_\epsilon]=\sum_{l=1}^\infty\mathbb{P}[|X_l-X|>\epsilon]<\infty$, implying that $N_\epsilon$ is finite a.s.

Now let us reason by contradiction and suppose that $\exists E$ with $\mathbb{P}[E]>0$ such that under $E$: $X_l \overset{l\to\infty}{\not\longrightarrow} X$. Under $E$, $\exists\epsilon$ random variable, and $\exists(k_l)_{l\in\mathbb{N}}$ random strictly increasing sequence such that $\forall l$: $|X_{k_l}-X|>\epsilon$. This implies in turn that $\exists E'$ with $\mathbb{P}[E']>0$ and $\exists\epsilon'>0$ non-random, such that under $E'$: $\exists(k_l)_{l\in\mathbb{N}}$ random strictly increasing sequence with $\forall l$: $|X_{k_l}-X|>\epsilon'$. Thus $N_{\epsilon'}$ 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 $A_l$, $A'_l$ of probability exponentially small in the width (see Section D.4 for justification);*

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

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

*Then, under $D$:*

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

**Proof.** Neglecting the events $A_l$, $A'_l$, Theorem 1 implies that $\exists m_l, m'_l, s_l, s'_l$ such that

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

On the other hand, under standard initialization:

$$\begin{aligned}\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]&=\mathbb{E}_{\Theta^{l-1}}\mathbb{E}_{\theta^l}\big[\nu_2(\mathbf{x}^{l-1})\cdot\delta\nu_2(\mathbf{x}^l)\big]=\mathbb{E}_{\Theta^{l-1}}\big[\nu_2(\mathbf{x}^{l-1})\cdot\mathbb{E}_{\theta^l}[\delta\nu_2(\mathbf{x}^l)]\big]=\mathbb{E}_{\Theta^{l-1}}\big[\nu_2(\mathbf{x}^{l-1})\big],\\ \mathbb{E}_{\Theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]&=\mathbb{E}_{\Theta^{l-1}}\mathbb{E}_{\theta^l}\big[\mu_2(\mathrm{d}\mathbf{x}^{l-1})\cdot\delta\mu_2(\mathrm{d}\mathbf{x}^l)\big]=\mathbb{E}_{\Theta^{l-1}}\big[\mu_2(\mathrm{d}\mathbf{x}^{l-1})\cdot\mathbb{E}_{\theta^l}[\delta\mu_2(\mathrm{d}\mathbf{x}^l)]\big]=\mathbb{E}_{\Theta^{l-1}}\big[\mu_2(\mathrm{d}\mathbf{x}^{l-1})\big],\end{aligned}$$

implying by induction that $\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)]=\nu_2(\mathbf{x}^0)$ and $\mathbb{E}_{\Theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]=\mu_2(\mathrm{d}\mathbf{x}^0)$.

Since $\log\nu_2(\mathbf{x}^l)$, $\log\mu_2(\mathrm{d}\mathbf{x}^l)$ are Gaussian by the assumption of lognormality, and since a logormal variable $\exp(X)$ with $X\sim\mathcal{N}(\mu,\sigma^2)$ has expectation equal to $\mathbb{E}[\exp(X)]=\exp(\mu+\sigma^2/2)$, it follows that $\exists S_l, S'_l$ random variables and $\exists M_l, M'_l>0$ constants such that

$$\begin{aligned}\log\nu_2(\mathbf{x}^l)&=S_l-M_l+\log\nu_2(\mathbf{x}^0), && S_l\sim_{\Theta^l}\mathcal{N}(0,2M_l), && lm_{\text{min}}\leq M_l\leq lm_{\text{max}},\\ \log\mu_2(\mathrm{d}\mathbf{x}^l)&=S'_l-M'_l+\log\mu_2(\mathrm{d}\mathbf{x}^0), && S'_l\sim_{\Theta^l}\mathcal{N}(0,2M'_l), && lm_{\text{min}}\leq M'_l\leq lm_{\text{max}}.\end{aligned}$$

Now let us make more precise the conditionality on $D$. We may assume that $\exists m>\frac{1}{2}\big(m_{\text{max}}-m_{\text{min}}\big)$ such that $\forall l$ under $D$: $\log\chi^l\geq lm$.

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

$$\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} = \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\mu_2(\mathbf{x}^0)}\frac{\mu_2(\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^l)}\right)\left(\frac{\mu_2(\mathbf{x}^0)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)}\right) = \frac{1}{(\chi^l)^2}\frac{\mu_2(\mathbf{x}^0)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)}.$$

This gives with logarithms that, under $D$:

$$\begin{aligned}\log\mu_2(\mathbf{x}^l) - \log\nu_2(\mathbf{x}^l) &= -2\log\chi^l + \log\mu_2(\mathrm{d}\mathbf{x}^l) - \log\mu_2(\mathrm{d}\mathbf{x}^0) - \log\nu_2(\mathbf{x}^l) + \log\mu_2(\mathbf{x}^0)\\ &\leq -2lm + \big(S'_l - M'_l\big) - \big(S_l - M_l + \log\nu_2(\mathbf{x}^0)\big) + \log\mu_2(\mathbf{x}^0)\\ &\leq -2lm + lm_{\text{max}} - lm_{\text{min}} - \log\nu_2(\mathbf{x}^0) + \log\mu_2(\mathbf{x}^0) + S'_l - S_l\\ &\leq -lM + C + S'_l - S_l,\end{aligned}$$

where we defined $M \equiv 2m - m_{\text{max}} + m_{\text{min}} > 0$ and $C \equiv -\log\nu_2(\mathbf{x}^0) + \log\mu_2(\mathbf{x}^0)$. Then for given $\epsilon$, under $D$:

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

where we denoted $\vee$ the logical *or*, $\tilde{S}_l \equiv S_l/\sqrt{2M_l}$ and $\tilde{S}'_l \equiv S'_l/\sqrt{2M'_l}$, and supposed $l$ large enough such that $\log\epsilon + lM - C \geq 0$. Then $\exists C_\epsilon > 0$ such that for $l$ large enough, under $D$:

$$\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon \implies \Big(\tilde{S}'_l \geq \sqrt{l}C_\epsilon\Big) \vee \Big(-\tilde{S}_l \geq \sqrt{l}C_\epsilon\Big).$$

It follows that for $l$ large enough:

$$\begin{aligned}\mathbb{P}_{\Theta^l|D}\left[\frac{\mu_2(\mathbf{x}^l)}{\nu_2(\mathbf{x}^l)} > \epsilon\right] &\leq \mathbb{P}_{\Theta^l|D}\Big[\tilde{S}'_l \geq \sqrt{l}C_\epsilon\Big] + \mathbb{P}_{\Theta^l|D}\Big[-\tilde{S}_l \geq \sqrt{l}C_\epsilon\Big]\\ &\leq \frac{1}{\mathbb{P}_{\Theta^l}[D]}\mathbb{P}_{\Theta^l}\Big[D \cap \big\{\tilde{S}'_l \geq \sqrt{l}C_\epsilon\big\}\Big] + \frac{1}{\mathbb{P}_{\Theta^l}[D]}\mathbb{P}_{\Theta^l}\Big[D \cap \big\{-\tilde{S}_l \geq \sqrt{l}C_\epsilon\big\}\Big]\\ &\leq \frac{1}{\mathbb{P}_{\Theta^l}[D]}\mathrm{erfc}\left(\sqrt{\frac{l}{2}}C_\epsilon\right) &(68)\\ &\leq \frac{1}{\mathbb{P}_{\Theta^l}[D]}\exp\Big(-\frac{l}{2}C_\epsilon^2\Big), &(69)\end{aligned}$$

where Eq. (68) is obtained using $\tilde{S}_l, \tilde{S}'_l \sim_{\Theta^l} \mathcal{N}(0,1)$, while Eq. (69) is obtained using $\mathrm{erfc}\,(x) \leq \exp(-x^2)$ (Chiani et al., 2003). It follows from Eq. (69) that

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

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

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

### D.8. If $\exp\big(\overline{m}[\chi^l]\big) \to 1$ and if Moments of $\tilde{\mathrm{x}}^l$ Are Bounded, then $\mathrm{x}^l$ Converges to One-Dimensional Signal Pathology

**Proposition 13.** *Again we adopt the notation: $\tilde{\mathbf{x}}^l \equiv \mathbf{x}^l / \sqrt{\mu_2(\mathbf{x}^l)}$, 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 $\nu_2(\mathbf{x}^l)/\mu_2(\mathbf{x}^l) \overset{l\to\infty}{\not\longrightarrow} \infty$ and thus $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \overset{l\to\infty}{\not\longrightarrow} 0$, 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:

$$\begin{aligned}
\boldsymbol{\nu}^l_\varphi &\equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})\big] = \big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{x}}^l)\big)_{1\leq \mathrm{c} \leq N_l}, \\
\boldsymbol{\nu}^l_\rho &\equiv \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^l,\boldsymbol{\alpha})\big].
\end{aligned}$$

The statistic-preserving property implies $\frac{1}{N_l}||\boldsymbol{\nu}^l_\varphi||_2^2 = \frac{1}{R_l}||\boldsymbol{\nu}^l_\rho||_2^2$, in turn implying that

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

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

*Now let us reason by contradiction and suppose that $r_{\text{eff}}(\mathbf{x}^l) = r_{\text{eff}}(\tilde{\mathbf{x}}^l) \overset{l\to\infty}{\not\longrightarrow} 1$, implying that $\exists\eta > 0$ and $\exists(k_l)_{l\in\mathbb{N}}$ strictly increasing sequence with $\forall l$: $r_{\text{eff}}(\tilde{\mathbf{x}}^{k_l}) \geq 1+\eta$.*

This directly implies that $\exists\eta' > 0$ such that $\forall l$:

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

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

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

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

$$\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:} = w_{\mathbf{v}}\big(\mathbf{v}^{k_l}_\rho\big)^T + \sqrt{1-w^2_{\mathbf{v}}}\,\mathbf{w}^T, \qquad \mathbf{w} \perp \mathbf{v}^{k_l}_\rho, \qquad ||\mathbf{w}|| = 1.$$

Then we get

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big] \\
&\quad= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[w_{\mathbf{v}}^2\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\rho^{k_l}\rangle^2 + (1-w_{\mathbf{v}}^2)\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle^2 + 2w_{\mathbf{v}}\sqrt{1-w_{\mathbf{v}}^2}\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}_\rho^{k_l}\rangle\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big] \\
&\quad\geq w_{\mathbf{v}}^2\eta' + 2w_{\mathbf{v}}\sqrt{1-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\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big] \\
&\quad\geq w_{\mathbf{v}}^2\eta' - 2w_{\mathbf{v}}\sqrt{1-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]^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle^2\Big]^{\frac{1}{2}} \\
&\quad\geq w_{\mathbf{v}}^2\eta' - 2w_{\mathbf{v}}\sqrt{1-w_{\mathbf{v}}^2}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\frac{\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})}{||\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})||}\right\rangle^2\right]^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\frac{\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})}{||\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})||}\right\rangle^2\right]^{\frac{1}{2}} \\
&\quad\geq w_{\mathbf{v}}^2\eta' - 2w_{\mathbf{v}}\sqrt{1-w_{\mathbf{v}}^2}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\textstyle\sum_i\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^2\Big] \\
&\quad\geq w_{\mathbf{v}}^2\eta' - 2w_{\mathbf{v}}\sqrt{1-w_{\mathbf{v}}^2}R_{k_l}\nu_2(\tilde{\mathbf{x}}^{k_l}), \\
&\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2 \\
&\quad= (1-w_{\mathbf{v}}^2)\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big]^2 \\
&\quad\leq (1-w_{\mathbf{v}}^2)\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\frac{\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})}{||\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})||}\right\rangle^2\right] \\
&\quad\leq (1-w_{\mathbf{v}}^2)R_{k_l}\nu_2(\tilde{\mathbf{x}}^{k_l}).
\end{aligned}$$

Given that $\nu_2(\tilde{\mathbf{x}}^{k_l})=\mathcal{O}(1)$, this implies by spherical symmetry that $\forall\epsilon>0$, $\exists p_\epsilon>0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\left[\left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2\geq\eta'^2-\epsilon\right)\wedge\left(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2\leq\epsilon\right)\right]\geq p_\epsilon, \tag{70}$$

with $\wedge$ the logical *and*.

On the other hand, by Cauchy-Schwarz inequality:

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2\leq\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big]\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|^3\Big]. \tag{71}$$

The second term on the right-hand side can be bounded as

$$\begin{aligned}
&\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|^3\Big] \\
&\qquad\leq\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\frac{\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})}{||\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})||_2}\right\rangle^3\right]=\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[||\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})||_2^3\Big]=\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\Big(\sum_{i=1}^{R_{k_l}}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^2\Big)^{3/2}\right] \\
&\qquad\leq\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\sum_{i_1,i_2,i_3}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_1}^2\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_2}^2\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_3}^2\right]^{1/2} &&(72)\\
&\qquad\leq\sum_{i_1,i_2,i_3}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_1}^4\big]^{1/4}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_2}^8\big]^{1/8}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{i_3}^8\big]^{1/8} &&(73)\\
&\qquad\leq R_{k_l}^3N_{k_l}^{1/2}\nu_4(\tilde{\mathbf{x}}^{k_l})^{1/4}\nu_8(\tilde{\mathbf{x}}^{k_l})^{1/4}, &&(74)
\end{aligned}$$

where Eq. (72) and Eq. (73) were obtained by applying Cauchy-Schwarz inequality, while Eq. (74) was obtained with $\forall i,\forall p$: $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^p\big]\leq\sum_{\mathrm{c}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\varphi(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_{\mathrm{c}}^p\big]=N_{k_l}\nu_p(\tilde{\mathbf{x}}^{k_l})$.

It then follows from Eq. (71) and the hypothesis that all moments are bounded $\nu_p(|\tilde{\mathbf{x}}^l|) = \mathcal{O}(1)$ that

$$\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^2\Big]^2 = \mathcal{O}\Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big]\Big). \tag{75}$$

Combining Eq. (70) and Eq. (75), we deduce that $\exists\eta'' > 0$ with $\forall\epsilon > 0$, $\exists p'_\epsilon > 0$ such that $\forall l$:

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

Under standard initialization: $\boldsymbol{W}^{k_l+1}_{\mathrm{c},:} \sim_{\theta^{k_l+1}} \mathcal{N}(0, 2\,/\,R_{k_l}\boldsymbol{I})$, the variables $\tilde{\boldsymbol{W}}^{k_l+1}_{\mathrm{c},:}$ and $||\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}||_2$ are independent and $\mathbb{P}_{\theta^{k_l+1}}\big[1 \leq ||\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}||_2 \leq 2\big] > 0$ does not depend on $l$. Therefore $\forall\epsilon > 0$, $\exists p''_\epsilon > 0$ such that $\forall l$:

$$\mathbb{P}_{\theta^{k_l+1}}\Bigg[\Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big] \geq \eta'' - \epsilon\Big) \wedge \Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2 \leq 4\epsilon\Big)\Bigg] \geq p''_\epsilon. \tag{76}$$

Now by noting that

$$\begin{aligned}
\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big|\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big|\Big] &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^+\Big] + \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^-\Big],\\
\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\Big]^2 &= \Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^+\Big] - \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^-\Big]\Big)^2,
\end{aligned}$$

we deduce that $\exists\eta''' > 0$, $\exists p > 0$ such that $\forall l$:

$$\begin{aligned}
\mathbb{P}_{\theta^{k_l+1}}\Bigg[\Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^+\Big] \geq \eta'''\Big) \wedge \Big(\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^{k_l+1}_{\mathrm{c},:}\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})\big)^-\Big] \geq \eta'''\Big)\Bigg] &\geq p,\\
\mathbb{P}_{\theta^{k_l+1}}\Bigg[\Big(\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,+}) \geq \eta'''\sqrt{\mu_2(\mathbf{x}^{k_l})}\Big) \wedge \Big(\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,-}) \geq \eta'''\sqrt{\mu_2(\mathbf{x}^{k_l})}\Big)\Bigg] &\geq p,\\
\mathbb{P}_{\theta^{k_l+1}}\Bigg[\frac{\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,-})}{\mu_2(\mathbf{x}^{k_l})} \geq (\eta''')^2\Bigg] &\geq p,\\
\mathbb{E}_{\mathrm{c},\theta^{k_l+1}}\Bigg[\frac{\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{k_l+1,-})}{\mu_2(\mathbf{x}^{k_l})}\Bigg] &\geq p(\eta''')^2.
\end{aligned}$$

Thus by Theorem 2, $\exists\eta'''' > 0$ such that $\forall l$: $\exp\big(\overline{m}[\chi^{k_l+1}]\big) \geq 1 + \eta''''$, contradicting the hypothesis $\exp\big(\overline{m}[\chi^l]\big) \xrightarrow{l\to\infty} 1$.

*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.* □

### D.9. If $\exp\big(\overline{m}[\chi^l]\big) \to 1$, then each Additional Layer $l$ Becomes Arbitrarily Well Approximated by a Linear Mapping

We suppose that $\forall l$: $\mu_2(\mathbf{x}^l) > 0$ and that $\exp\big(\overline{m}[\chi^l]\big) \to 1$. Denoting $\tilde{\mathbf{y}}^l = \mathbf{y}^l/\sqrt{\mu_2(\mathbf{x}^{l-1})}$ and $\tilde{\mathbf{y}}^{l,\pm} \equiv \max\big(\pm\tilde{\mathbf{y}}^l, 0\big)$, Theorem 2 implies that

$$\begin{aligned}
\mathbb{E}_{\mathrm{c},\theta^l}\big[\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big] &\to 0,\\
\mathbb{E}_{\mathrm{c},\theta^l}\Big[\min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big)^2\Big] &\to 0,\\
\forall\epsilon > 0: \quad \mathbb{P}_{\mathrm{c},\theta^l}\big[\min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big) > \epsilon\big] &\to 0,\\
\forall\epsilon > 0: \quad \mathbb{P}_{\theta^l}\big[\exists\mathrm{c}:\ \min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big) > \epsilon\big] &\to 0,\\
\forall\epsilon > 0: \quad \mathbb{P}_{\theta^l}\big[\forall\mathrm{c}:\ \min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+}), \nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big) \leq \epsilon\big] &\to 1.
\end{aligned} \tag{77}$$

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

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

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 incurred when approximating the rescaled signal $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1}) = \mathbf{y}^{l,+}/\mu_2(\mathbf{x}^{l-1}) = \tilde{\mathbf{y}}^{l,+}$ in channel c by a linear function. 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

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

If $\forall$c: $\min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big) \leq \epsilon$, 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

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

Combined with Eq. (77), this means that $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1}) = \tilde{\mathbf{y}}^{l,+}$ can be approximated arbitrarily well by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrarily close to 1 in $\theta^l$.

We have shown that $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1})$ is arbitrarily well approximated by a linear function of $\mathbf{x}^{l-1}$ when normalizing with respect to $\mathbf{x}^{l-1}$. Now let us show that $\tilde{\mathbf{x}}^l = \mathbf{x}^l/\mu_2(\mathbf{x}^l)$ is arbitrarily well approximated by a linear function of $\mathbf{x}^{l-1}$ when normalizing with respect to $\mathbf{x}^l$.

Let us denote $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_{R_l})$ and $(\lambda_1,\dots,\lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^l \equiv \boldsymbol{W}^l(\boldsymbol{e}_1,\dots,\boldsymbol{e}_{R_l})$. By Corollary 3 there is at least one eigenvalue $\lambda_i$ such that $\lambda_i \geq \mu_2(\mathbf{x}^{l-1})$, which gives combined with Eq. (60) that $\forall$c:

$$\mu_{2,\mathrm{c}}(\tilde{\mathbf{y}}^l) = \frac{1}{\mu_2(\mathbf{x}^{l-1})}\sum\nolimits_i \big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2\lambda_i,$$
$$\mu_{2,\mathrm{c}}(\tilde{\mathbf{y}}^l) \geq X, \qquad X \sim_{\theta^l} \frac{2}{R_l}\,\text{Chi-Squared}(1).$$

Using Eq. (65), we then get

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = \mu_{2,\mathrm{c}}(\mathbf{y}^l) - 2\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-}) = \mu_2(\mathbf{x}^{l-1})\Big(\mu_{2,\mathrm{c}}(\tilde{\mathbf{y}}^l) - 2\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\Big),$$
$$\mu_{2,\mathrm{c}}(\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) \geq \mu_2(\mathbf{x}^{l-1})\big(X - Y\big), \qquad X \sim_{\theta^l} \frac{2}{R_l}\,\text{Chi-Squared}(1), \qquad \forall\epsilon : \mathbb{P}_{\theta^l}\big[|Y| > \epsilon\big] \xrightarrow{l\to\infty} 0. \tag{78}$$

Similarly to the proof of Theorem 1, we define

$$w^l_\mathrm{c} \equiv \begin{cases} 0 & \text{if } \mu_{2,\mathrm{c}}(\mathbf{x}^l) < \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) \\ 1 & \text{if } \mu_{2,\mathrm{c}}(\mathbf{x}^l) > \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) \\ \tilde{w}^l_\mathrm{c} & \text{if } \mu_{2,\mathrm{c}}(\mathbf{x}^l) = \mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) \end{cases},$$

with $\tilde{w}^l_\mathrm{c} \sim$ Bernouilli$(1/2)$ independent of $\boldsymbol{\omega}^l$ and $\boldsymbol{\beta}^l$.

Since $C_l$ is independent from $\mu_{2,\mathrm{c}}(\mathbf{x}^l)+\mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)$, it follows from Eq. (78) that $\forall p > 0, \exists\eta, \eta' > 0$ such that for $l$ large enough:

$$\mathbb{P}_{\theta^l|C_l}\left[\mu_2(\mathbf{x}^l) \geq \frac{1}{2N_l}\mu_2(\mathbf{x}^{l-1})\eta\right] > 1 - p, \tag{79}$$

$$\mathbb{P}_{\theta^l|C_l}\left[\frac{\mu_2(\mathbf{x}^l)}{\mu_2(\mathbf{x}^{l-1})} \geq \eta'\right] > 1 - p. \tag{80}$$

Now let us fix $p, \epsilon > 0$ and consider $\eta'$ as in Eq. (80). If we suppose that $\forall$c: $\min\big(\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,+})\nu_{1,\mathrm{c}}(\tilde{\mathbf{y}}^{l,-})\big) \leq \sqrt{\eta'}\epsilon$, and that $\frac{\mu_2(\mathbf{x}^l)}{\mu_2(\mathbf{x}^{l-1})} \geq \eta'$, then there exists a 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

$$\begin{aligned}\nu_1(|\tilde{\mathbf{y}}^{l,+} - f(\mathbf{x}^{l-1})|) &\leq \sqrt{\eta'}\epsilon,\\ \nu_1(|\tilde{\mathbf{x}}^{l} - \tilde{f}(\mathbf{x}^{l-1})|) &\leq \sqrt{\frac{\mu_2(\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^l)}}\sqrt{\eta'}\epsilon \leq \frac{1}{\sqrt{\eta'}}\sqrt{\eta'}\epsilon = \epsilon,\end{aligned}$$

where we defined $\tilde{f}(\mathbf{x}^{l-1}) = \sqrt{\frac{\mu_2(\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^l)}} f(\mathbf{x}^{l-1})$. Given Eq. (77), this means that $\tilde{\mathbf{x}}^l$ can be approximated with error $\epsilon$ by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrarily close to $(1-p)\mathbb{P}_{\theta^l}[C_l] = (1-p)\big(1-2^{-N_l}\big)$. Thus $\tilde{\mathbf{x}}^l$ can be approximated arbitrarily well by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrarily close to $\mathbb{P}_{\theta^l}[C_l] = 1-2^{-N_l}$. Furthermore $\mathbb{P}_{\theta^l}[C_l]$ is itself nearly indistinguishable from 1.

## E. Details of Section 6

### E.1. Proof of Theorem 3

**Theorem 3** (normalized sensitivity increments of batch-normalized feedforward nets)**.** *The dominating term in the evolution of $\chi^l$ can be decomposed as*

$$\begin{aligned}\delta\chi^l = \delta_{\mathrm{BN}}\chi^l \,\cdot\, \delta_\phi\chi^l \simeq \exp\big(\overline{m}[\chi^l]\big) &= \exp\big(\overline{m}_{\mathrm{BN}}[\chi^l]\big) \,\cdot\, \exp\big(\overline{m}_\phi[\chi^l]\big),\\ \exp\big(\overline{m}_{\mathrm{BN}}[\chi^l]\big) &\equiv \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-\frac{1}{2}} \mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right]^{\frac{1}{2}},\\ \exp\big(\overline{m}_\phi[\chi^l]\big) &\equiv \underbrace{\Big(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})]\Big)^{-\frac{1}{2}}}_{\in[1,\sqrt{2}]}.\end{aligned}$$

**Proof.** First let us decompose $\delta\chi^l$ as the product of $\delta_{\mathrm{BN}}\chi^l$ and $\delta_\phi\chi^l$:

$$\begin{aligned}\delta_{\mathrm{BN}}\chi^l &\equiv \left(\frac{\mu_2(\mathrm{d}\mathbf{z}^l)}{\mu_2(\mathbf{z}^l)}\right)^{\frac{1}{2}}\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-\frac{1}{2}},\\ \delta_\phi\chi^l &\equiv \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathbf{x}^l)}\right)^{\frac{1}{2}}\left(\frac{\mu_2(\mathrm{d}\mathbf{z}^l)}{\mu_2(\mathbf{z}^l)}\right)^{-\frac{1}{2}},\\ \delta\chi^l &= \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathbf{x}^l)}\right)^{\frac{1}{2}}\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l-1})}{\mu_2(\mathbf{x}^{l-1})}\right)^{-\frac{1}{2}} = \delta_{\mathrm{BN}}\chi^l \,\cdot\, \delta_\phi\chi^l.\end{aligned}$$

Next let us decompose $\exp\big(\overline{m}[\chi^l]\big)$ as the product of two terms:

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

The term $\exp\big(\overline{m}_{\mathrm{BN}}[\chi^l]\big)$ approximates the geometric increment $\delta_{\mathrm{BN}}\chi^l$ from $(\mathbf{x}^{l-1},\mathrm{d}\mathbf{x}^{l-1})$ to $(\mathbf{z}^l,\mathrm{d}\mathbf{z}^l)$ such that $\exp\big(\overline{m}_{\mathrm{BN}}[\chi^l]\big)\simeq\delta_{\mathrm{BN}}\chi^l$, while the term $\exp\big(\overline{m}_{\phi}[\chi^l]\big)$ approximates the geometric increment $\delta_\phi\chi^l$ from $(\mathbf{z}^l,\mathrm{d}\mathbf{z}^l)$ to $(\mathbf{x}^l,\mathrm{d}\mathbf{x}^l)$ such that $\exp\big(\overline{m}_{\phi}[\chi^l]\big)\simeq\delta_\phi\chi^l$. These terms can be seen (slightly simplistically) as the direct contribution of respectively batch normalization and the nonlinearity $\phi$ to $\delta\chi^l$. Now let us explicitate both terms.

**Term** $\exp\big(\overline{m}_{\mathrm{BN}}\big[\chi^l\big]\big)$**.** First let us note that batch normalization directly gives $\mu_2(\mathbf{z}^l)=1$, and thus $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]=1$. Next let us explicitate $\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{z}^l)]$:

$$\forall\mathrm{c}:\ \mathrm{d}\mathbf{z}^l_{:,\mathrm{c}}=\frac{\mathrm{d}\mathbf{y}^l_{:,\mathrm{c}}}{\sqrt{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}},\qquad\forall\mathrm{c}:\ \mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{z}^l)=\frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)},$$

$$\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{z}^l)]=\mathbb{E}_{\mathrm{c},\theta^l}[\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{z}^l)]=\mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^l)}{\mu_{2,\mathrm{c}}(\mathbf{y}^l)}\right].$$

All together, we get that

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

**Term** $\exp\big(\overline{m}_{\phi}\big[\chi^l\big]\big)$**.** We consider the symmetric propagation for batch-normalized feedforward nets, introduced in Section B. From Eq. (26), we deduce that

$$\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]+\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\bar{\mathbf{x}}^l)]=\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{z}^l)],$$

$$2\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]=\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{z}^l)],\tag{81}$$

where Eq. (81) is obtained by symmetry of the propagation. Next we turn to the symmetric propagation of the signal:

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{x}^l)+\mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)&=\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[(\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}})^2\big]-\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{z}^{l,+}_{\boldsymbol{\alpha},\mathrm{c}}\big]^2+\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[(\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}})^2\big]-\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathbf{z}^{l,-}_{\boldsymbol{\alpha},\mathrm{c}}\big]^2 \qquad (82)\\&=\nu_{2,\mathrm{c}}(\mathbf{z}^{l,+})-\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})^2+\nu_{2,\mathrm{c}}(\mathbf{z}^{l,-})-\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})^2\\&=\nu_{2,\mathrm{c}}(\mathbf{z}^l)-\Big(\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})^2+\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})^2\Big),\end{aligned}$$

where Eq. (82) follows from Eq. (23). Due to the constraints $\nu_{1,\mathrm{c}}(\mathbf{z}^l)=0$ and $\nu_{2,\mathrm{c}}(\mathbf{z}^l)=1$, imposed by batch normalization:

$$\mu_{2,\mathrm{c}}(\mathbf{x}^l)+\mu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)=1-\Big(\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})^2+\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})^2\Big),\tag{83}$$

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

$$\big(\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})-\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})\big)^2=\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})^2+\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})^2-2\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})=0.\tag{84}$$

Using Eq. (83), Eq. (84) and the symmetry of the propagation:

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

$$2\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)]=1-2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,-}\big)].\tag{85}$$

Finally combining Eq. (81), Eq. (85) and $\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]=1$:

$$\begin{aligned}\exp\big(\overline{m}_{\phi}\big[\chi^l\big]\big)&=\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{x}^l)]}\right)^{\frac{1}{2}}\left(\frac{\mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{z}^l)]}{\mathbb{E}_{\theta^l}[\mu_2(\mathbf{z}^l)]}\right)^{-\frac{1}{2}}\\&=\Big(1-2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,+}\big)\nu_{1,\mathrm{c}}\big(\mathbf{z}^{l,-}\big)]\Big)^{-\frac{1}{2}}.\end{aligned}$$

To obtain the bounds on $\exp\big(\overline{m}_\phi[\chi^l]\big)$, the same reasoning as Eq. (67) may be applied to $\mathbf{z}^l$ instead of $\mathbf{y}^l$:

$$4\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-}) \leq \mu_{2,\mathrm{c}}(\mathbf{z}^l) = 1, \qquad 2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})] \leq \frac{1}{2},$$

$$1 \leq \exp\big(\overline{m}_\phi[\chi^l]\big) = \Big(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})]\Big)^{-\frac{1}{2}} \leq \sqrt{2}. \qquad \square$$

### E.2. In the First Step of the Propagation, $\exp\big(\overline{m}_{\mathrm{BN}}[\chi^1]\big) \geq 1$

Using again the notations from Section B, we may explicitate the second-order moment in channel c of $\mathrm{d}\mathbf{y}^1$:

$$\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^1) = \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\Big[\hat{\varphi}(\mathrm{d}\mathbf{y}^1,\boldsymbol{\alpha})_\mathrm{c}^2\Big] = \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\Big[\varphi(\mathrm{d}\mathbf{y}^1,\boldsymbol{\alpha})_\mathrm{c}^2\Big] = \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\Big[\big(\boldsymbol{W}^1_{\mathrm{c},:}\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})\big)^2\Big] \tag{86}$$

$$= \sum\nolimits_{i,j} \boldsymbol{W}^1_{\mathrm{c},i}\boldsymbol{W}^1_{\mathrm{c},j}\mathbb{E}_{\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_i\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_j]$$

$$= \mu_2(\mathrm{d}\mathbf{x}^0)\sum\nolimits_i \big(\boldsymbol{W}^1_{\mathrm{c},i}\big)^2 = \mu_2(\mathrm{d}\mathbf{x}^0)\,||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2. \tag{87}$$

where Eq. (86) follows from $\mathrm{d}\mathbf{y}^1$ being centered, while Eq. (87) follows from the white noise property $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i\mathrm{d}\mathbf{x}_j] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x}^0)\,\delta_{ij}$, implying $\forall\boldsymbol{\alpha}$: $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_i\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_j] = \mu_2(\mathrm{d}\mathbf{x}^0)\,\delta_{ij}$ 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 $\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x},\boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^1 = \boldsymbol{W}^1(\boldsymbol{e}_1,\dots,\boldsymbol{e}_{R_1})$, we get that

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

$$= ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2 \sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i = \frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^1)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i, \tag{88}$$

where we defined $\tilde{\boldsymbol{W}}^1$ such that $\forall$c: $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:} = \hat{\boldsymbol{W}}^1_{\mathrm{c},:} \,/\, ||\hat{\boldsymbol{W}}^1_{\mathrm{c},:}||$ and we used Eq. (87). Under standard initialization, the distribution of $\hat{\boldsymbol{W}}^1$ is spherically symmetric, implying that for all channels c the distribution of $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:}$ is uniform on the unit sphere of $\mathbb{R}^{R_1}$. In turn, this implies that

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

$$\forall \mathrm{c}:\ \mathbb{E}_{\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big] = \frac{1}{R_1}\sum\nolimits_i \lambda_i, \qquad \mathbb{E}_{\mathrm{c},\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big] = \frac{1}{R_1}\sum\nolimits_i \lambda_i. \tag{89}$$

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

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

$$= \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\frac{1}{R_1}\sum_i\lambda_i}\right)^{-\frac{1}{2}} \mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\sum_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i}\right]^{\frac{1}{2}} \tag{90}$$

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

$$\geq \left(\frac{1}{R_1}\sum_i \lambda_i\right)^{\frac{1}{2}} \left(\mathbb{E}_{\mathrm{c},\theta^1}\Big[\sum\nolimits_i \big(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\big)^2\lambda_i\Big]^{-1}\right)^{\frac{1}{2}} = 1. \tag{91}$$

where Eq. (90) was obtained using Eq. (88) and $\mu_2(\mathbf{x}^0) = \mu_2(\mathbf{x}) = \frac{1}{R_1}\operatorname{Tr}\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x},\boldsymbol{\alpha})] = \frac{1}{R_1}\sum_i\lambda_i$ by Corollary 3, while Eq. (91) was obtained using the convexity of $x \mapsto 1/x$ and Eq. (89).

# F. Details of Section 7

## F.1. Adaptation of the Previous Setup to Resnets

Before proceeding to the analysis, slight adaptations and forewords are necessary. We denote

$$\Theta^{l,h} \equiv (\boldsymbol{\omega}^{1,1}, \boldsymbol{\beta}^{1,1}, \dots, \boldsymbol{\omega}^{1,H}, \boldsymbol{\beta}^{1,H}, \dots, \boldsymbol{\omega}^{l,1}, \boldsymbol{\beta}^{l,1}, \dots, \boldsymbol{\omega}^{l,h}, \boldsymbol{\beta}^{l,h}), \qquad \theta^{l,h} \equiv \Theta^{l,h} | \Theta^{l,h-1},$$
$$\Theta^{l} \equiv (\boldsymbol{\omega}^{1,1}, \boldsymbol{\beta}^{1,1}, \dots, \boldsymbol{\omega}^{1,H}, \boldsymbol{\beta}^{1,H}, \dots, \boldsymbol{\omega}^{l,1}, \boldsymbol{\beta}^{l,1}, \dots, \boldsymbol{\omega}^{l,H}, \boldsymbol{\beta}^{l,H}), \qquad \theta^{l} \equiv \Theta^{l} | \Theta^{l-1}.$$

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

Since the evolution of $\chi^l$ is mainly influenced by batch normalization and the nonlinearity $\phi$, 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 $(\mathbf{x}^{l,h-1}, \mathrm{d}\mathbf{x}^{l,h-1})$ to $(\mathbf{x}^{l,h}, \mathrm{d}\mathbf{x}^{l,h})$ for layers $2 \leq h \leq H$. Theorem 3 strictly applies in this case.

It remains to understand the evolution from $(\mathbf{y}^{l,0}, \mathrm{d}\mathbf{y}^{l,0}) = (\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1})$ to $(\mathbf{x}^{l,1}, \mathrm{d}\mathbf{x}^{l,1})$ in layer $h = 1$, and the evolution from $(\mathbf{x}^{l,H}, \mathrm{d}\mathbf{x}^{l,H})$ to $(\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H})$ in layer $h = H$.

By considering the parameter $\Theta^{l-1}$, the dominating term in the evolution from $(\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1})$ to $(\mathbf{z}^{l,1}, \mathrm{d}\mathbf{z}^{l,1})$ is

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

Under the assumption of well-conditioned noise, this term is again $\gtrsim 1$ by convexity of $x \mapsto 1/x$. For the nonlinearity term, the symmetric propagation with respect to $\Theta^{l-1}$ applies for all terms in the sum $(\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1}) = \sum_{k=0}^{l-1} (\mathbf{y}^{k,H}, \mathrm{d}\mathbf{y}^{k,H})$, except for $(\mathbf{y}^{0,H}, \mathrm{d}\mathbf{y}^{0,H}) = (\mathbf{y}, \mathrm{d}\mathbf{y})$. The expression of the nonlinearity term $\exp\big(\overline{m}_\phi[\chi^l]\big)$ in Theorem 3 thus remains approximately valid.

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

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

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

## F.2. Lemma on Dot-Products

**Lemma 14.** *It holds that:*

$$\begin{aligned}
\mathbb{E}_{\theta^l}\big[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathbf{y}^{l-1})_\mathrm{c}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big]\big] &= 0, \\
\mathbb{E}_{\theta^l}\Big[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathbf{y}^{l-1})_\mathrm{c}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big]^2\Big] &\leq \frac{1}{N r_{\text{eff}}(\mathbf{y}^{l-1})} \mu_2(\mathbf{y}^{l-1}) \mathbb{E}_{\theta^l}[\mu_2(\mathbf{y}^{l,H})], \\
\mathbb{E}_{\theta^l}\big[\mathbb{E}_{\mathbf{y},\mathrm{d}\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathrm{d}\mathbf{y}^{l-1})_\mathrm{c}\hat{\varphi}(\mathrm{d}\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big]\big] &= 0, \\
\mathbb{E}_{\theta^l}\Big[\mathbb{E}_{\mathbf{y},\mathrm{d}\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathrm{d}\mathbf{y}^{l-1})_\mathrm{c}\hat{\varphi}(\mathrm{d}\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big]^2\Big] &\leq \frac{1}{N r_{\text{eff}}(\mathrm{d}\mathbf{y}^{l-1})} \mu_2(\mathrm{d}\mathbf{y}^{l-1}) \mathbb{E}_{\theta^l}[\mu_2(\mathrm{d}\mathbf{y}^{l,H})].
\end{aligned}$$

**Proof.** By spherical symmetry, the 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 $\theta^l$.

It follows that

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

Next we note that

$$\begin{aligned}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\big[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_\mathrm{c}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big] &= \frac{1}{N}\sum\nolimits_\mathrm{c}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_\mathrm{c}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\big]\\ &= \frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha}),\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big], \qquad (92)\end{aligned}$$

with $\langle\,,\rangle$ 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 $\boldsymbol{C}_{\mathbf{y},\boldsymbol{\alpha}}[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})]$. Let us 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 $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$. Then we get that

$$\begin{gathered}\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha}) = \sum\nolimits_i\sqrt{\lambda_i}u_i\boldsymbol{e}_i, \qquad \forall i:\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[u_i^2\big]=1, \qquad \forall j\neq i:\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[u_iu_j\big]=0,\\ \hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha}) = \sum\nolimits_i y_i\boldsymbol{e}_i.\end{gathered}$$

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

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

From this definition, we get that

$$\begin{gathered}\forall j:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_iu_j\big]=0, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_iu_i\big]=\alpha_{i,i}, \qquad \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_i^2\big]=\sum\nolimits_j\alpha_{i,j}^2+\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_i^2\big],\\ \mu_2(\mathbf{y}^{l,H})=\frac{1}{N}\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\Big[\big\langle\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha}),\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\big\rangle\Big]=\frac{1}{N}\sum\nolimits_i\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_i^2\big]=\frac{1}{N}\Big(\sum\nolimits_{i,j}\alpha_{i,j}^2+\sum\nolimits_i\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[z_i^2\big]\Big), \qquad (93)\end{gathered}$$

where the dot product in Eq. (93) was 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 $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_N)$:

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

Spherical symmetry implies that the moments of $y_1\boldsymbol{e}_1+\dots+y_i\boldsymbol{e}_i+\dots+y_N\boldsymbol{e}_N$ and $y_1\boldsymbol{e}_1+\dots-y_i\boldsymbol{e}_i+\dots+y_N\boldsymbol{e}_N$ have the same distribution with respect to $\theta^l$. Thus $\forall j\neq i$:

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

We deduce that

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

Spherical symmetry also implies that the distribution of $\alpha_{i,j}$ with respect to $\theta^l$ does not depend on $i$. Denoting $(\beta_j)$ such that $\forall i, j$: $\beta_j \equiv \mathbb{E}_{\theta^l}\left[\alpha_{i,j}^2\right]$, we get combined with Eq. (93):

$$\begin{aligned}
\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right] &\geq \frac{1}{N}\sum\nolimits_{i,j}\mathbb{E}_{\theta^l}\left[\alpha_{i,j}^2\right] = \frac{1}{N}\sum\nolimits_{i,j}\beta_j \geq \sum\nolimits_i \beta_i,\\
\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right]^2\right] &= \sum\nolimits_i \lambda_i\beta_i \leq \lambda_{\text{max}}\Big(\sum\nolimits_i \beta_i\Big)\\
&\leq \lambda_{\text{max}}\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right].
\end{aligned}$$

Finally combining with Eq. (92):

$$\begin{aligned}
\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\right]^2\right] &= \frac{1}{N^2}\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\left[\left\langle\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})\right\rangle\right]^2\right]\\
&\leq \frac{1}{N^2}\lambda_{\text{max}}\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right] \qquad (94)\\
&\leq \frac{1}{N r_{\text{eff}}(\mathbf{y}^{l-1})}\mu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\mu_2(\mathbf{y}^{l,H})\right],
\end{aligned}$$

where we used $\lambda_{\text{max}} r_{\text{eff}}(\mathbf{y}^{l-1}) = \sum_i \lambda_i = N\mu_2(\mathbf{y}^{l-1})$.

The same analysis immediately applies to $\hat{\varphi}(\mathrm{d}\mathbf{y}^{l-1},\boldsymbol{\alpha})$ and $\hat{\varphi}(\mathrm{d}\mathbf{y}^{l,H},\boldsymbol{\alpha})$. □

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

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

*Then by spherical symmetry $\forall l$, $\forall l' \neq l$:*

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

*Furthermore given Lemma 14 and given $r_{\text{eff}}(\mathbf{y}^{l-1})$, $r_{\text{eff}}(\mathrm{d}\mathbf{y}^{l-1}) \geq 1$, we deduce that*

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

*These inequalities will be useful in the proof of Theorem 4.*

### F.3. Proof of Theorem 4

**Theorem 4** (normalized sensitivity increments of batch-normalized resnets)**.** *Suppose that we can bound signal variances: $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ and feedforward increments: $\delta_{\min} \lesssim \delta\chi^{l,h} \lesssim \delta_{\max}$ for all $l,h$. Further denote $\eta_{\min} \equiv \big((\delta_{\min})^{2H}\mu_{2,\min} - \mu_{2,\max}\big)/\mu_{2,\max}$ and $\eta_{\max} \equiv \big((\delta_{\max})^{2H}\mu_{2,\max} - \mu_{2,\min}\big)/\mu_{2,\min}$, as well as $\tau_{\min} \equiv \eta_{\min}/2$ and $\tau_{\max} \equiv \eta_{\max}/2$. Then there exist positive constants $C_{\min}$, $C_{\max} > 0$ such that*

$$\Big(1+\frac{\eta_{\min}}{l+1}\Big)^{\frac{1}{2}} \lesssim \delta\chi^l \lesssim \Big(1+\frac{\eta_{\max}}{l+1}\Big)^{\frac{1}{2}},$$
$$C_{\min} l^{\tau_{\min}} \lesssim \ \chi^l \ \lesssim C_{\max} l^{\tau_{\max}}.$$

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

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:

$$\begin{aligned}
(a \lesssim b) \wedge (b \lesssim a) &\iff (a \simeq b),\\
(a \lesssim b) &\iff (-a \gtrsim -b),\\
(a \lesssim b) \wedge (b \lesssim c) &\implies (a \lesssim c),\\
(a \lesssim b) \wedge (c \lesssim d) \wedge (a>0) \wedge (c>0) &\implies (ac \lesssim bd),\\
(a \lesssim b) \wedge (c \lesssim d) \wedge (a>0) \wedge (c>0) &\implies (a+c \lesssim b+d),\\
(a \lesssim b) \wedge (a>0) \wedge (b>0) &\implies (\sqrt{a} \lesssim \sqrt{b}),\\
(a \lesssim b) \wedge (a>0) \wedge (b>0) &\implies (1/a \gtrsim 1/b).
\end{aligned}$$

Finally let $a$ be a random variable depending on $\Theta^l$ with well-defined moments and let $b$ be a constant. Let us prove that

$$(a \lesssim b) \implies \big(\mathbb{E}_{\theta^l}[a] \lesssim b\big) \wedge \big(\mathbb{E}_{\Theta^l}[a] \lesssim b\big).$$

Given the assumption $(a \lesssim b)$, there exists an event $A$ with $\mathbb{P}_{\Theta^l}[A] \simeq 1$ such that under $A$: $a(1+\epsilon_a) \leq b(1+\epsilon_b)$ with $|\epsilon_a| \ll 1$, $|\epsilon_b| \ll 1$. Furthermore, using Cauchy-Schwarz inequality:

$$\frac{1}{\mathbb{E}_{\theta^l}[a]^2}\Big(\mathbb{E}_{\theta^l}[a] - \mathbb{E}_{\theta^l}[\mathbf{1}_A a]\Big)^2 = \frac{1}{\mathbb{E}_{\theta^l}[a]^2}\mathbb{E}_{\theta^l}[\mathbf{1}_{A^c} a]^2 \leq \mathbb{P}_{\theta^l}[A^c]\frac{\mathbb{E}_{\theta^l}[a^2]}{\mathbb{E}_{\theta^l}[a]^2}. \tag{95}$$

Since $\mathbb{P}_{\Theta^l}[A] \simeq 1$, the complementary event $A^c$ has probability $\mathbb{P}_{\Theta^l}[A^c] \ll 1$. Now by contradiction, if there would be non negligible probability with respect to $\Theta^{l-1}$ that $\mathbb{P}_{\theta^l}[A^c] = \mathbb{P}_{\Theta^l|\Theta^{l-1}}[A^c]$ is non negligible, then we would not have that $\mathbb{P}_{\Theta^l}[A^c] = \mathbb{E}_{\Theta^{l-1}}\mathbb{E}_{\Theta^l|\Theta^{l-1}}[\mathbf{1}_{A^c}] = \mathbb{E}_{\Theta^{l-1}}\mathbb{P}_{\Theta^l|\Theta^{l-1}}[A^c]$ is negligible. It follows that $\mathbb{P}_{\theta^l}[A^c] \ll 1$ with high probability with respect to $\Theta^{l-1}$.

Combined with Eq. (95) and the definition of $A$, we get

$$\mathbb{E}_{\theta^l}[a] \simeq \mathbb{E}_{\theta^l}[\mathbf{1}_A a] \lesssim b.$$

A similar reasoning gives

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

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

**Proof of Eq. (14).** Adopting the notations of Corollary 15 and using $\mathbf{y}^l=\mathbf{y}^{l-1}+\mathbf{y}^{l,H}$ by Eq. (12), we get that

$$\mu_2(\mathbf{y}^l)=\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\Big(\hat{\varphi}(\mathbf{y}^{l-1},\boldsymbol{\alpha})_\mathrm{c}+\hat{\varphi}(\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\Big)^2\right]=\mu_2(\mathbf{y}^{l-1})+Y_{l,l}+2Y_l, \tag{96}$$

$$\mu_2(\mathrm{d}\mathbf{y}^l)=\mathbb{E}_{\mathbf{y},\mathrm{d}\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\left[\Big(\hat{\varphi}(\mathrm{d}\mathbf{y}^{l-1},\boldsymbol{\alpha})_\mathrm{c}+\hat{\varphi}(\mathrm{d}\mathbf{y}^{l,H},\boldsymbol{\alpha})_\mathrm{c}\Big)^2\right]=\mu_2(\mathrm{d}\mathbf{y}^{l-1})+T_{l,l}+2T_l.$$

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

Now let us reason by induction and suppose that $l\mu_{2,\text{min}}\lesssim\mu_2(\mathbf{y}^{l-1})\lesssim l\mu_{2,\text{max}}$. Combined with Eq. (96), we get that

$$l\mu_{2,\text{min}}+\mu_{2,\text{min}}+2Y_l\lesssim\mu_2(\mathbf{y}^l)\lesssim l\mu_{2,\text{max}}+\mu_{2,\text{max}}+2Y_l.$$

On the other hand, Corollary 15 implies that

$$\mathbb{E}_{\Theta^l}\left[Y_l^2\right]\lesssim\frac{1}{N}l\mu_{2,\text{max}}^2\leq\frac{1}{N}\frac{1}{l+1}(l+1)^2\mu_{2,\text{max}}^2.$$

Further using Chebyshev's inequality, we deduce that

$$\mathbb{P}_{\Theta^l}\left[\big|Y_l\big|>k\frac{1}{\sqrt{N}}\frac{1}{\sqrt{l+1}}(l+1)\mu_{2,\text{max}}\right]\lesssim\frac{1}{k^2}. \tag{97}$$

For large width $N\gg1$, it follows that $\big|Y_l\big|\ll(l+1)\mu_{2,\text{min}}$ and $\big|Y_l\big|\ll(l+1)\mu_{2,\text{max}}$ with high probability, and thus that

$$(l+1)\mu_{2,\text{min}}\lesssim\mu_2(\mathbf{y}^l)\lesssim(l+1)\mu_{2,\text{max}}. \tag{98}$$

Then Eq. (98) holds for all $l$, and furthermore $\big|Y_l\big|\ll\mu_2(\mathbf{y}^{l-1})$ with high probability. Now let us write $(\chi^l)^2$ as

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

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

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

$$(\delta\chi^l)^2=\frac{(\chi^l)^2}{(\chi^{l-1})^2}=\frac{\mu_2(\mathbf{y}^{l-1})+\tilde{T}_{l,l}+2\tilde{T}_l}{\mu_2(\mathbf{y}^{l-1})+Y_{l,l}+2Y_l}. \tag{99}$$

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

$$\tilde{T}_{l,l}=\frac{\mu_2(\mathbf{y}^0)}{\mu_2(\mathrm{d}\mathbf{y}^0)(\chi^{l-1})^2}\mu_2(\mathrm{d}\mathbf{y}^{l,H})=\frac{\mu_2(\mathbf{y}^0)}{\mu_2(\mathrm{d}\mathbf{y}^0)(\chi^{l-1})^2}(\chi^{l-1})^2\prod_h(\delta\chi^{l,h})^2\mu_2(\mathbf{y}^{l,H})\frac{\mu_2(\mathrm{d}\mathbf{y}^0)}{\mu_2(\mathbf{y}^0)},$$

$$\gamma_\text{min}\mu_{2,\text{min}}\lesssim\tilde{T}_{l,l}\lesssim\gamma_\text{max}\mu_{2,\text{max}}. \tag{100}$$

By Corollary 15, the variance of $\tilde{T}_l$ is bounded as

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

The same reasoning as Eq. (97) implies both $\big|Y_l\big| \ll \mu_2(\mathbf{y}^{l-1})$ and $\big|\tilde{T}_l\big| \ll \mu_2(\mathbf{y}^{l-1})$ with high probability. Finally combining Eq. (99), Eq. (100) and the hypothesis $\mu_{2,\text{min}} \lesssim Y_{l,l} \lesssim \mu_{2,\text{max}}$:

$$\begin{aligned}
\frac{\mu_2(\mathbf{y}^{l-1})+\gamma_{\text{min}}\mu_{2,\text{min}}}{\mu_2(\mathbf{y}^{l-1})+\mu_{2,\text{max}}} \lesssim (\delta\chi^l)^2 &\lesssim \frac{\mu_2(\mathbf{y}^{l-1})+\gamma_{\text{max}}\mu_{2,\text{max}}}{\mu_2(\mathbf{y}^{l-1})+\mu_{2,\text{min}}}, \\
1+\frac{\gamma_{\text{min}}\mu_{2,\text{min}}-\mu_{2,\text{max}}}{\mu_2(\mathbf{y}^{l-1})+\mu_{2,\text{max}}} \lesssim (\delta\chi^l)^2 &\lesssim 1+\frac{\gamma_{\text{max}}\mu_{2,\text{max}}-\mu_{2,\text{min}}}{\mu_2(\mathbf{y}^{l-1})+\mu_{2,\text{min}}}, \\
1+\frac{\gamma_{\text{min}}\mu_{2,\text{min}}-\mu_{2,\text{max}}}{(l+1)\mu_{2,\text{max}}} \lesssim (\delta\chi^l)^2 &\lesssim 1+\frac{\gamma_{\text{max}}\mu_{2,\text{max}}-\mu_{2,\text{min}}}{(l+1)\mu_{2,\text{min}}}, \\
\Big(1+\frac{\eta_{\text{min}}}{l+1}\Big)^{\frac{1}{2}} \lesssim \quad \delta\chi^l \quad &\lesssim \Big(1+\frac{\eta_{\text{max}}}{l+1}\Big)^{\frac{1}{2}},
\end{aligned}$$

where we supposed $\big(\gamma_{\text{min}}\mu_{2,\text{min}}-\mu_{2,\text{max}}\big) \geq 0$ (see Section F.1 and the evolution of Fig. 4 for the justification). □

**Proof of Eq. (15).** Expanding Eq. (14), we get that

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

We can further explicitate the bounds:

$$\begin{aligned}
&\sum_{k=1}^{l}\log\Big(1+\frac{\eta_{\text{max}}}{k+1}\Big) \\
&\qquad \leq \int_1^{l+1}\log\Big(1+\frac{\eta_{\text{max}}}{x}\Big)dx \\
&\qquad \leq \int_1^{l+1}\log(x+\eta_{\text{max}})dx - \int_1^{l+1}\log x\,dx \\
&\qquad \leq \Big[x\log x - x\Big]_{1+\eta_{\text{max}}}^{l+1+\eta_{\text{max}}} - \Big[x\log x - x\Big]_1^{l+1} \\
&\qquad \leq (l+1+\eta_{\text{max}})\log(l+1+\eta_{\text{max}}) - (1+\eta_{\text{max}})\log(1+\eta_{\text{max}}) - (l+1)\log(l+1) \\
&\qquad \leq \eta_{\text{max}}\log(l+1+\eta_{\text{max}}) + (l+1)\log\Big(1+\frac{\eta_{\text{max}}}{l+1}\Big) - (1+\eta_{\text{max}})\log(1+\eta_{\text{max}}) \\
&\qquad \leq \eta_{\text{max}}\log(l+1+\eta_{\text{max}}) + \eta_{\text{max}} - (1+\eta_{\text{max}})\log(1+\eta_{\text{max}}), \qquad (101)
\end{aligned}$$

where we used $\log(1+x) \leq x$ in Eq. (101). Considering the integration between $2$ and $l+2$, we similarly get:

$$\begin{aligned}
&\sum_{k=1}^{l}\log\Big(1+\frac{\eta_{\text{min}}}{k+1}\Big) \\
&\qquad \geq \eta_{\text{min}}\log(l+2+\eta_{\text{min}}) + (l+2)\log\Big(1+\frac{\eta_{\text{min}}}{l+2}\Big) - (2+\eta_{\text{min}})\log(2+\eta_{\text{min}}) + 2\log 2 \\
&\qquad \geq \eta_{\text{min}}\log(l+2+\eta_{\text{min}}) - (2+\eta_{\text{min}})\log(2+\eta_{\text{min}}) + 2\log 2.
\end{aligned}$$

Let $c_{\text{max}} \equiv \exp\Big(\eta_{\text{max}} - (1+\eta_{\text{max}})\log(1+\eta_{\text{max}})\Big)$ and $c_{\text{min}} \equiv \exp\Big(-(2+\eta_{\text{min}})\log(2+\eta_{\text{min}}) + 2\log 2\Big)$. Then:

$$\prod_{k=1}^{l}\Big(1+\frac{\eta_{\text{max}}}{k+1}\Big) \leq c_{\text{max}}(l+1+\eta_{\text{max}})^{\eta_{\text{max}}}, \qquad \prod_{k=1}^{l}\Big(1+\frac{\eta_{\text{min}}}{k+1}\Big) \geq c_{\text{min}}(l+2+\eta_{\text{min}})^{\eta_{\text{min}}},$$

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

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

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

### F.4. Theorem 4 Holds for any Choice of $\phi$, with and without Batch Normalization, as long as the Existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$, $\delta_{\text{min}}$, $\delta_{\text{max}}$ is Ensured

The proof of Lemma 14 neither requires batch normalization nor does it require any assumption on $\phi$. In addition, the proof still holds up to Eq. (94) when replacing $\hat{\varphi}(\mathbf{y}^{l-1})$, $\hat{\varphi}(\mathbf{y}^{l,H})$, $\mu_2(\mathbf{y}^{l-1})$, $\mu_2(\mathbf{y}^{l,H})$ by $\varphi(\mathbf{y}^{l-1})$, $\varphi(\mathbf{y}^{l,H})$, $\nu_2(\mathbf{y}^{l-1})$, $\nu_2(\mathbf{y}^{l,H})$ and eigenvalues of $\boldsymbol{C}_{\mathbf{y},\boldsymbol{\alpha}}[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})]$ by eigenvalues of $\boldsymbol{G}_{\mathbf{y},\boldsymbol{\alpha}}[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})]$. This gives

$$\begin{aligned}\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\Big[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\varphi(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\Big]^2\right] &\leq \frac{1}{N^2}\lambda_{\text{max}}\mathbb{E}_{\theta^l}\left[\nu_2(\mathbf{y}^{l,H})\right] \\ &\leq \frac{1}{N}\nu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\nu_2(\mathbf{y}^{l,H})\right]. \qquad (102)\end{aligned}$$

Similarly, the proof of Theorem 4 only depends on batch normalization and the choice of $\phi$ through the constants $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$, $\delta_{\text{min}}$, $\delta_{\text{max}}$. As a consequence, Theorem 4 holds for any choice of $\phi$, with and without batch normalization, as long as the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$, $\delta_{\text{min}}$, $\delta_{\text{max}}$ is ensured.

It is therefore interesting to determine in which cases the constants $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$, $\delta_{\text{min}}$, $\delta_{\text{max}}$ exist. In the forthcoming analysis, we will consider the common cases $\phi = \tanh$ and $\phi = \mathrm{ReLU}$, with and without batch normalization, relating our results and providing extensions to Yang & Schoenholz (2017).

*For the sake of brevity, some results will be established only with an informal proof.*

#### F.4.1. Case $\phi = \tanh$, without Batch Normalization

From $\mathbf{x}^{l,H} = \phi(\mathbf{y}^{l,H-1})$, we deduce that $\nu_2(\mathbf{x}^{l,H})$, $\mu_2(\mathbf{x}^{l,H})$ are bounded as

$$\mu_2(\mathbf{x}^{l,H}) \leq \nu_2(\mathbf{x}^{l,H}) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\phi(\mathbf{y}^{l,H-1})^2\right] \leq 1.$$

Since $\mathbf{y}^{l,H}$ is obtained from $\mathbf{x}^{l,H}$ only after a single single convolution step, it follows that $\nu_2(\mathbf{y}^{l,H})$, $\mu_2(\mathbf{y}^{l,H})$ are bounded from above. Let us further admit that $\nu_2(\mathbf{y}^{l,H})$, $\mu_2(\mathbf{y}^{l,H})$ are bounded from below so that the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$ is ensured.

Now let us see whether $\delta_{\text{min}}$, $\delta_{\text{max}}$ exist in the mean-field limit: $N \to \infty$, where $\mathbf{y}^l$ becomes a Gaussian process and all moment-related quantities become deterministic with the expectation over $\Theta^l$ equivalent to the average over channels. Using Lemma 14 as well as Eq. (102), combined with the reasoning of Eq. (98) on $\nu_2(\mathbf{y}^l)$ and $\mu_2(\mathbf{y}^l)$ for large $N \gg 1$:

$$\nu_2(\mathbf{y}^{l-1}) \propto l, \qquad \mu_2(\mathbf{y}^{l-1}) \propto l.$$

The probability of non-negligible $\phi'(\mathbf{y}^{l,0})^2 = \phi'(\mathbf{y}^{l-1})^2$ is equal to the probability that $\mathbf{y}^{l-1}$ is roughly $\mathcal{O}(1)$, which scales as $\frac{1}{\sqrt{\nu_2(\mathbf{y}^{l-1})}}\frac{1}{\sqrt{2\pi}} \propto \frac{1}{\sqrt{l}}$ for large $l$. Combined with $\mathrm{d}\mathbf{x}^{l,1} = \phi'(\mathbf{y}^{l,0}) \odot \mathrm{d}\mathbf{y}^{l,0}$, this implies that

$$\frac{\mu_2(\mathrm{d}\mathbf{x}^{l,1})}{\mu_2(\mathrm{d}\mathbf{y}^{l,0})} \propto \frac{1}{\sqrt{l}}.$$

Given $\mu_2(\mathbf{x}^{l,1}) \leq \nu_2(\mathbf{x}^{l,1}) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\phi(\mathbf{y}^{l,0})^2\right] \leq 1$, we get for the ratio of signal variances:

$$\frac{\mu_2(\mathbf{y}^{l,0})}{\mu_2(\mathbf{x}^{l,1})} \geq \mu_2(\mathbf{y}^{l-1}) \propto l.$$

This gives for the *squared* geometric increment during the nonlinearity step from $(\mathbf{y}^{l,0}, \mathrm{d}\mathbf{y}^{l,0})$ to $(\mathbf{x}^{l,1}, \mathrm{d}\mathbf{x}^{l,1})$:

$$\left(\frac{\mu_2(\mathrm{d}\mathbf{x}^{l,1})}{\mu_2(\mathbf{x}^{l,1})}\right)\left(\frac{\mu_2(\mathrm{d}\mathbf{y}^{l,0})}{\mu_2(\mathbf{y}^{l,0})}\right)^{-1} \geq \mu_2(\mathbf{y}^{l-1})\frac{\mu_2(\mathrm{d}\mathbf{x}^{l,1})}{\mu_2(\mathrm{d}\mathbf{y}^{l,0})} \propto \sqrt{l}.$$

It follows that $\delta\chi^{l,1}$ and thus $\delta\chi^{l,h}$ are *not* bounded from above and that the existence of $\delta_{\text{max}}$ is *not* ensured. Now if we replace $\eta_{\text{min}}$, $\eta_{\text{max}}$ by $\frac{\mathcal{A}}{2}\sqrt{l+1} \propto \sqrt{l}$ in Eq. (14):

$$\delta\chi^l \simeq \left(1 + \frac{\mathcal{A}}{2\sqrt{l+1}}\right)^{\frac{1}{2}}.$$

Given $\frac{1}{2}\log(1 + \frac{\mathcal{A}}{2\sqrt{x}}) \simeq \frac{\mathcal{A}}{4\sqrt{x}}$ and $\int_{x_0}^{x} \frac{\mathcal{A}}{4\sqrt{x'}}dx' \simeq \frac{\mathcal{A}}{2}\sqrt{x}$ for $x \gg 1$, we get that $\chi^l = \prod_k \delta\chi^k = \exp\big(\sum_k \log\delta\chi^l\big) \propto \exp\big(\frac{\mathcal{A}}{2}\sqrt{l}\big)$. Combined with $\mu_2(\mathbf{y}^{l-1}) \propto l$ and the definition of $\chi^l$, we deduce that $\mu_2(\mathrm{d}\mathbf{y}^{l-1}) \propto \exp\big(\mathcal{A}\sqrt{l}\big)$, which is exactly the scaling found in Yang & Schoenholz (2017) for the corresponding quantity.

*In summary, the growth of $\chi^l$ is slightly subexponential but still far from power-law.*

### F.4.2. Case $\phi = \tanh$, with Batch Normalization

Batch normalization controls signal variance inside residual units: $\mu_2(\mathbf{z}^{l,H}) = 1$. Since $\mathbf{y}^{l,H}$ is obtained from $\mathbf{z}^{l,H}$ only after a single nonlinearity step and a single convolution step, the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$ is ensured.

Now let us see whether $\delta_{\text{min}}$, $\delta_{\text{max}}$ exist and let us first limit our reasoning to the feedforward evolution of Section 6. Since the reasoning of Section 6 on the effect of batch normalization applies for any choice of $\phi$, the assumption of well-conditioned noise implies that $\exp(\overline{m}_{\mathrm{BN}}[\chi^l])$ is bounded: (i) from above by considering the signal with worst possible conditioning; (ii) from below by 1.

Regarding $\exp(\overline{m}_\phi[\chi^l])$, let us consider again the mean-field limit: $N \to \infty$ such that $\mathbf{z}^l$ is Gaussian with variance equal to $\nu_2(\mathbf{z}^l) = \mu_2(\mathbf{z}^l) = 1$. Then $\exp(\overline{m}_\phi[\chi^l])$ is deterministic and constant, implying that $\exp(\overline{m}_\phi[\chi^l])$ is bounded by constants from above and below.

Since the evolution inside residual units is well approximated by the feedforward evolution of Section 6, it follows that $\delta\chi^{l,h}$ is bounded from above and below.

*In summary, Theorem 4 applies and $\chi^l$ has power-law growth.*

### F.4.3. Case $\phi = \mathrm{ReLU}$, without Batch Normalization

Since the evolution inside residual units is well approximated by the feedforward evolution of Section 5, it follows that $\nu_2(\mathbf{y}^{l,h})$, $\mu_2(\mathrm{d}\mathbf{y}^{l,h})$ are roughly stable and that the increments $\delta\chi^{l,h}$ are limited inside residual units. This implies that $\nu_2(\mathbf{y}^{l,H}) \simeq \nu_2(\mathbf{y}^{l,0}) = \nu_2(\mathbf{y}^{l-1})$ and $\mu_2(\mathbf{y}^{l,H}) \simeq \mu_2(\mathbf{y}^{l,0}) = \mu_2(\mathbf{y}^{l-1})$. Combined with Eq. (96) and the fact that $Y_{l,l} = \mu_2(\mathbf{y}^{l,H})$ and that $Y_l \ll \mu_2(\mathbf{y}^{l-1})$ with high probability for $N \gg 1$, we deduce that

$$\mu_2(\mathbf{y}^l) \simeq \mu_2(\mathbf{y}^{l-1}) + \mu_2(\mathbf{y}^{l,H}) \simeq 2\mu_2(\mathbf{y}^{l-1}).$$

Using Eq. (102), the same reasoning for non-central moments gives

$$\nu_2(\mathbf{y}^l) \simeq \nu_2(\mathbf{y}^{l-1}) + \nu_2(\mathbf{y}^{l,H}) \simeq 2\nu_2(\mathbf{y}^{l-1}).$$

This means that both $\mu_2(\mathbf{y}^l)$ and $\nu_2(\mathbf{y}^l)$ have exponential growth and that the existence of $\mu_{2,\text{max}}$ is not ensured. The exponential growth of $\nu_2(\mathbf{y}^l)$ agrees with the scaling found for the corresponding quantity in Yang & Schoenholz (2017).

Now let us see whether $\delta_{\text{min}}$, $\delta_{\text{max}}$ exist. In the feedforward evolution of Section 5, Theorem 2 directly ensures that $1 \lesssim \delta\chi^l \lesssim \sqrt{2}$.

Again since the evolution inside residual units is well approximated by the feedforward evolution of Section 5, we deduce that $\delta\chi^{l,h}$ is bounded from above and below.

*In summary, the existence of $\mu_{2,\text{max}}$ is not ensured and Theorem 4 does not apply. No conclusion can be made regarding the growth of $\chi^l$.*

### F.4.4. CASE $\phi = \mathrm{ReLU}$, WITH BATCH NORMALIZATION

As in Section F.4.2, the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$ is ensured by the fact that batch normalization controls signal variance: $\mu_2(\mathbf{z}^{l,H}) = 1$ and that $\mathbf{y}^{l,H}$ is obtained from $\mathbf{z}^{l,H}$ only after a single nonlinearity step and a single convolution step.

Now let us see whether $\delta_\text{min}$, $\delta_\text{max}$ exist and again let us first reason in the feedforward evolution of Section 6. Similarly to Section F.4.2, the term $\exp(\overline{m}_\mathrm{BN}[\chi^l])$ is bounded: (i) from above by considering the signal with worst possible conditioning; (ii) from below by $1$.

Theorem 3 further ensures that $1 \leq \exp(\overline{m}_\phi[\chi^l]) \leq \sqrt{2}$.

Since the evolution inside residual units is well approximated by the feedforward evolution of Section 6, we deduce that $\delta\chi^{l,h}$ is bounded from above and below.

*In summary, Theorem 4 applies and $\chi^l$ has power-law growth.*