Characterizing Well-behaved vs. Pathological Deep Neural Networks

Antoine Labatie

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 the commonly used techniques of batch normalization and skip-connections. Our key insight is to consider the evolution with depth of statistical moments of signal and noise, thereby characterizing the presence or the absence of pathologies in the hypothesis space encoded by the choice of hyperparameters. We establish: (i) for feedforward networks with and without batch normalization, depth multiplicativity 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]

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 gradients (Yang & Schoenholz, 2017; Philipp et al., 2018; Yang et al., 2019). 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), "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 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 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 – collapsing data representations, exploding sensitivity – in feedforward networks. And 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.

Correspondence to: <Antoine Labatie antoine.labatie@centraliens.net>.

[1]Code to reproduce all results in this paper is 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 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 ϕ.[2] 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 tensors of 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 weight tensors and $\mathbf{b}^l \in \mathbb{R}^{N_l}$ the biases. Later in our analysis, the model parameters $\boldsymbol{\omega}^l$ and $\mathbf{b}^l$ will be considered as random but for now they are considered as *fixed*. The propagation at each layer is given by

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

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 $\mathrm{d}\mathbf{x} \in \mathbb{R}^{n\times\cdots\times n\times N_0}$ with small i.i.d. components such that $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i \mathrm{d}\mathbf{x}_j] = \sigma_{\mathrm{d}\mathbf{x}}^2 \delta_{ij}$, with $\sigma_{\mathrm{d}\mathbf{x}} \ll 1$ and δ_{ij} the Kronecker delta for multidimensional indices i, j. The noisy signal is propagated into the same neural network and we keep track of the noise corruption with the tensor $\mathrm{d}\mathbf{x}^l$:

$$\mathrm{d}\mathbf{x}^0 \equiv \mathrm{d}\mathbf{x}, \qquad \mathrm{d}\mathbf{x}^l \equiv \Phi_l(\mathbf{x} + \mathrm{d}\mathbf{x}) - \Phi_l(\mathbf{x}),$$

with Φ_l the neural network mapping from layer 0 to l such that $\mathbf{x}^l = \Phi_l(\mathbf{x})$. The simultaneous propagation of the signal $\mathbf{x}^l$ and the noise $\mathrm{d}\mathbf{x}^l$ is given by

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

where $\odot$ is the element-wise tensor multiplication and the propagation of $\mathrm{d}\mathbf{x}^l$ is obtained by differentiating the propagation of $\mathbf{x}^l$. Note 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 $\mathrm{d}\mathbf{x}$ 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 α and the channel as c.

[2] It is possible to relax the assumptions of periodic boundary conditions and fixed 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.

Scope. We require two mild assumptions: (i) $\mathbf{x}$ is not trivially zero: $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha},\mathrm{c}}[\mathbf{x}^2_{\boldsymbol{\alpha},\mathrm{c}}] > 0$;[3] (ii) the width N_l is bounded.

Some results of our analysis will apply for any choice of ϕ, but unless otherwise stated we assume 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.1].

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

3. Data randomness

We may now 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: $\mathbf{x}^L = \Phi_{l,L}(\mathbf{x}^l)$ and $\mathbf{x}^L + \mathrm{d}\mathbf{x}^L = \Phi_{l,L}(\mathbf{x}^l + \mathrm{d}\mathbf{x}^l)$, with $\Phi_{l,L}$ the *upper neural network* mapping 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, i.e. $P_{\mathbf{x},\boldsymbol{\alpha}}(\mathbf{x}^l)$, $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}(\mathrm{d}\mathbf{x}^l)$ 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 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*,

$$\varphi(\mathbf{v}^l, \boldsymbol{\alpha}) \equiv \mathbf{v}^l_{\boldsymbol{\alpha},:}, \qquad \hat{\varphi}(\mathbf{v}^l, \boldsymbol{\alpha}) \equiv \mathbf{v}^l_{\boldsymbol{\alpha},:} - \mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\mathbf{v}^l_{\boldsymbol{\alpha},:}].$$[4]

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

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

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

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

where $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}$ is the covariance matrix and $||\cdot||$ is the spectral norm. If we further denote (λ_i) the eigenvalues

[3] 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 $\{1, \ldots, N_l\}$.

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

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 χ^l, we may consider the signal-to-noise ratio 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 $\mathrm{SNR}^l_{\mathrm{db}} = \mathrm{SNR}^0_{\mathrm{db}} - 20\log_{10}\chi^l$ in logarithmic decibel scale, meaning that χ^l measures how the neural network from layer 0 to l degrades ($\chi^l > 1$) or enhances ($\chi^l < 1$) the input signal-to-noise ratio. Neural networks with $\chi^l \gg 1$ are *noise amplifiers* while neural networks with $\chi^l \ll 1$ are *noise reducers*.

Let us now justify our choice of terminology 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 χ^l exactly measures the *excess root mean square sensitivity* of the neural network mapping Φ_l relative to the constant rescaling Ψ_l [C.2]. This is illustrated in Fig. 1.

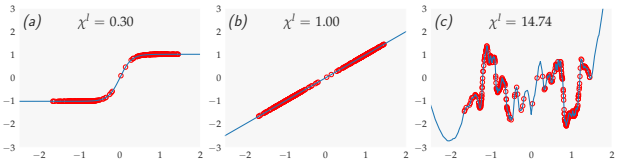

Figure 1: *Illustration of* χ^l in the fully-connected case with one-dimensional input and output: $N_0 = 1$ and $N_l = 1$. We show the full input-output mapping Φ_l *(blue curves)* and randomly sampled input-output data points $\left(\mathbf{x}^0, \Phi_l(\mathbf{x}^0)\right)$ *(red circles)* for three 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 Φ_l, the sensitivity is low: $\chi^l < 1$. *(b)* Φ_l is a constant rescaling: $\chi^l = 1$. *(c)* Since Φ_l is highly chaotic, the sensitivity is high: $\chi^l > 1$.

As just outlined, χ^l measures 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).[5] 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:

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

Then $||\tilde{\boldsymbol{\nu}}^l||_2 = 1$ and 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 point $\tilde{\boldsymbol{\nu}}^l$ [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 only one 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 Eq. (4), the pathology $\chi^l \to \infty$ implies: $\mathrm{SNR}^l \to 0$, meaning that the clean signal $\mathbf{x}^l$ becomes *drowned* in the noise $\mathrm{d}\mathbf{x}^l$.[6] 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 $\boldsymbol{\omega}^l \sim \mathcal{N}\left(0, 2\,/\,(K_l^d N_{l-1})\,\boldsymbol{I}\right)$,

[5] The coefficient defined in Philipp & Carbonell (2018) is equivalent to χ^l in the fully-connected case [C.3].

[6] In this case, the propagation of $\mathrm{d}\mathbf{x}^l$ eventually does not follow Eq. (1) and Eq. (2) since $\mathrm{d}\mathbf{x}^l$ becomes non-infinitesimal even for infinitesimal $\mathrm{d}\mathbf{x}$.

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 first justification, the distribution on input-output mappings at initialization can be seen as the prior encoded by the choice of hyperparameters in the context of Bayesian neural networks (Neal, 1996; Williams, 1997).

In the standard context of non-Bayesian neural networks, the randomness of model parameters still makes networks at initialization excellent proxy for the full hypothesis space. In fact, characterizing ReLU feedforward networks at initialization in terms of the pathologies of Section 3.2 is *strictly equivalent to characterizing the full hypothesis space restricted to zero biases and zero batch normalization shift*. To see this, notice that (i) a constant rescaling at a given layer only implies a constant rescaling at subsequent layers;[7] (ii) all pathologies of Section 3.2 are invariant to constant rescalings. Together (i) and (ii) imply that L^2 norms of the vectorized weights $||\operatorname{vec}(\boldsymbol{\omega}^l)||_2$ do not play any role and that any spherically symmetric initialization characterizes the full weight space. In other words, *a given probability of pathology at initialization corresponds to the same fraction of weight space associated with the pathology*.

As second justification, it is likely that pathologies at initialization would cause optimization issues irrespective of whether well-behaved input-output mappings exist in the hypothesis space or not. Let us illustrate this argument:

– 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 ϕ (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 $\phi = \mathrm{ReLU}$ or bounded ϕ (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).

[7] In the case of vanilla nets, this comes from the positive homogeneity property: $\Phi_{l,L}(r\mathbf{x}) = r\Phi_{l,L}(\mathbf{x})$, where $r \geq 0$ and $\Phi_{l,L}$ is the neural network mapping from layer $l < L$ to L with zero biases. In the case of batch-normalized feedforward nets, this comes from the fact that any constant rescaling during convolution is "undone" during batch normalization.

From now on, our methodology is to consider all moment-related quantities, e.g. $\nu_p(\mathbf{x}^l)$, $\mu_p(\mathbf{x}^l)$, $\mu_p(\mathrm{d}\mathbf{x}^l)$, $r_{\text{eff}}(\mathbf{x}^l)$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^l)$, χ^l, as random variables which depend on model parameters. We denote model parameters as $\Theta^l \equiv (\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \ldots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ and we use θ^l as shorthand for $\Theta^l|\Theta^{l-1}$. We further denote geometric increments 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 we expressed $\log\delta\nu_2(\mathbf{x}^k)$ with telescoping terms. Denoting the multiplicatively centered increments: $\underline{\delta}\nu_2(\mathbf{x}^k) \equiv \delta\nu_2(\mathbf{x}^k)/\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$, 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 Θ^{k-1} while $\underline{s}[\nu_2(\mathbf{x}^k)]$ is a random variable which depends on Θ^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 δ. 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 χ^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 that this assumption is milder than the *mean-field* assumption of infinite width: $N_l \to \infty$. Indeed mean-field would consider $\overline{m}[\chi^l]$ as deterministic and $\mathbf{y}^l$ as a Gaussian process indexed by $\mathbf{x}, \boldsymbol{\alpha}$, whereas we still consider $\overline{m}[\chi^l]$ as random and $\mathbf{y}^l$ as possibly a non-Gaussian process.

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* $1 \gg m_{\min}, m_{\max}, v_{\min}, v_{\max} > 0$*, 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*

$$\text{Under } A_l: \quad \log \nu_2(\mathbf{x}^l) = -l m_l + \sqrt{l} s_l + \log \nu_2(\mathbf{x}^0),$$
$$\text{Under } A'_l: \quad \log \mu_2(\mathrm{d}\mathbf{x}^l) = -l m'_l + \sqrt{l} s'_l + \log \mu_2(\mathrm{d}\mathbf{x}^0).$$

$$m_{\min} \leq m_l \leq m_{\max}, \quad \mathbb{E}_{\Theta^l|A_l}[s_l] = 0, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l|A_l}[s_l] \leq v_{\max}$$
$$m_{\min} \leq m'_l \leq m_{\max}, \quad \mathbb{E}_{\Theta^l|A'_l}[s'_l] = 0, \quad v_{\min} \leq \mathrm{Var}_{\Theta^l|A'_l}[s'_l] \leq v_{\max}$$

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 probability of the complementary events A_l^c, $A_l'^c$ decays 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].[8] 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 expression 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 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 only stabilized in terms of expectation and they become fat-tailed distributed as $l \to \infty$.

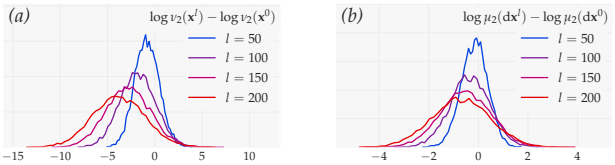

Figure 2: *Slowly diffusing moments of vanilla nets* with $L = 200$ layers of width $N_l = 128$. *(a)* Distribution of $\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)$ for $l = 50, 100, 150, 200$. *(b)* Same for $\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)$.

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

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

Discussion. A first consequence is that χ^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 γ 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 limiting case: $\delta\chi^l \simeq \exp\left(\overline{m}[\chi^l]\right) \to 1$, if the moments of the rescaled signal $\tilde{\mathbf{x}}^l \equiv \mathbf{x}^l/\sqrt{\mu_2(\mathbf{x}^l)}$ are bounded, then Theorem 2 implies the convergence to the pathology of one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$ [D.8] and the convergence to neural network *pseudo-linearity*, with each additional layer l becoming arbitrary 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: $\delta\chi^l \simeq \exp\left(\overline{m}[\chi^l]\right) \to 1$, 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.

[8] Any non-negligible difference with the scaling 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 $(\mathbf{x}^l, \mathrm{d}\mathbf{x}^l)$.

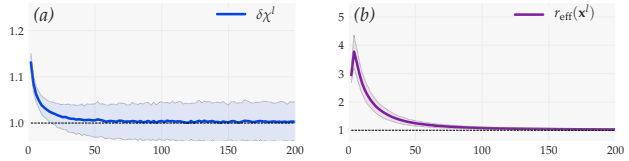

Figure 3: *Pathology of one-dimensional signal for vanilla nets* with $L = 200$ layers of width $N_l = 512$. *(a)* $\delta\chi^l$ such that $\delta\chi^l \simeq \exp\left(\overline{m}[\chi^l]\right) \to 1$. *(b)* $r_{\text{eff}}(\mathbf{x}^l)$ indicates one-dimensional signal pathology: $r_{\text{eff}}(\mathbf{x}^l)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

The only way that the neural network can achieve pseudo-linearity is by having each one of its 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* χ^l *can be decomposed as*

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

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

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

Effect of batch normalization. The batch normalization term satisfies: $\exp(\overline{m}_{\text{BN}}[\chi^l]) \simeq \delta_{\text{BN}}\chi^l$, with $\delta_{\text{BN}}\chi^l$ defined as the increment of χ^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 ϕ.

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 χ^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: $W_r^{l,c} \equiv \left\{||\mathrm{vec}(\boldsymbol{\omega}^l_{:,:,\text{c}})||_2 = r\right\}$, 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 expectations under $W_r^{l,c}$:

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

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 whether $r_{\text{eff}}(\mathbf{x}^{l-1})$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^{l-1})$ are small compared to the ambient space dimension N_{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 $x \mapsto 1/x$:

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

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 where 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 satisfies: $\exp(\overline{m}_\phi[\chi^l]) \simeq \delta_\phi\chi^l$, with $\delta_\phi\chi^l$ defined as the increment of χ^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 $\mathrm{d}\mathbf{x}^l$ remains well-conditioned while $\mathbf{x}^l$ becomes ill-conditioned. And that $r_{\text{eff}}(\mathbf{x}^l)$ 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 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 $|\delta_\phi\chi^l - 1|$.

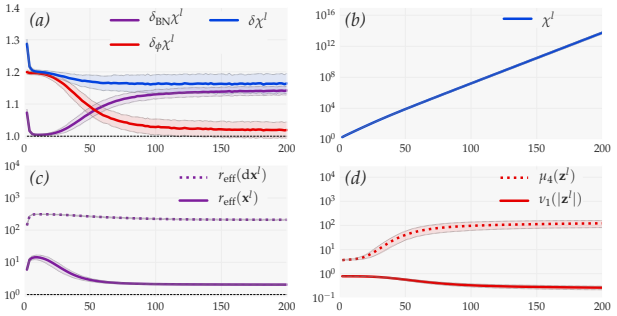

Figure 4: *Pathology of exploding sensitivity for batch-normalized feedforward nets* with $L = 200$ layers of width $N_l = 512$. *(a)* Geometric increments $\delta\chi^l$ decomposed as the product of $\delta_{\text{BN}}\chi^l$ defined as the increment from $(\mathbf{x}^{l-1}, \mathrm{d}\mathbf{x}^{l-1})$ 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 χ^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 $r_{\text{eff}}(\mathbf{x}^l) \ll N_l$. *(d)* $\mathbf{z}^l$ becomes fat-tailed distributed with respect to $\mathbf{x}$, $\boldsymbol{\alpha}$ with large $\mu_4(\mathbf{z}^l)$ and small $\nu_1(|\mathbf{z}^l|)$.

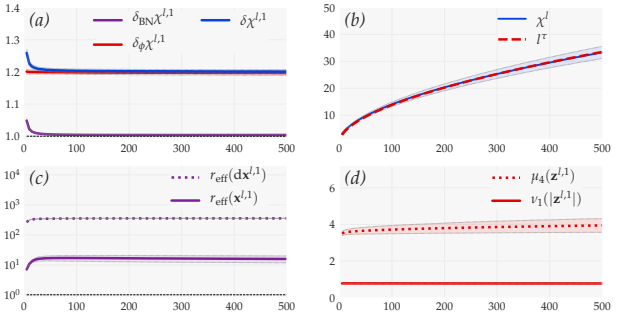

Figure 5: *Well-behaved evolution of batch-normalized resnets* with $L = 500$ residual units comprised of $H = 2$ layers of width $N = 512$. *(a)* Geometric feedforward increments $\delta\chi^{l,1}$ decomposed as the product of $\delta_{\text{BN}}\chi^{l,1}$ defined as the increment from $(\mathbf{y}^{l,0}, \mathrm{d}\mathbf{y}^{l,0})$ 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)* χ^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

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

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

$1 \leq h \leq H$, with H the number of layers inside residual units
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

$$(\mathbf{y}^l, \mathrm{d}\mathbf{y}^l) = \sum_{k=0}^{l} (\mathbf{y}^{k,H}, \mathrm{d}\mathbf{y}^{k,H}). \qquad (13)$$

For consistency reasons, we redefine the inputs of the propagation as $(\mathbf{y}^0, \mathrm{d}\mathbf{y}^0) \equiv (\mathbf{y}, \mathrm{d}\mathbf{y})$ 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:* $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ *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* $\tau_{\min} \equiv \frac{1}{2}\eta_{\min}$ *and* $\tau_{\max} \equiv \frac{1}{2}\eta_{\max}$. *Then there exist positive constants* $C_{\min}, C_{\max} > 0$ *such that*

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

$$C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}. \qquad (15)$$

Discussion. First let us note that Theorem 4 remarkably holds for any choice of ϕ with and without batch normalization, as long as the existence of $\mu_{2,\min}$, $\mu_{2,\max}$, $\delta_{\min}$, $\delta_{\max}$ is ensured. In the case $\phi = \mathrm{ReLU}$, the existence of $\delta_{\min}$, $\delta_{\max}$ is always ensured but the existence of $\mu_{2,\min}$, $\mu_{2,\max}$ is only ensured when batch normalization controls signal variance inside residual units: $\mu_2(\mathbf{z}^{l,H}) = 1$ [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 $\eta_{\min}$, $\eta_{\max}$ have an exponential dependence in H. However, it is slowed down by the factor $1/(l+1)$ between successive residual units. This is due to 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 set $\mu_{2,\min} = \mu_{2,\max}$ and if we remove the dilution effect by multiplying the residual branch

by 0, thus replacing the scaling in $1/(l+1)$ by a scaling in 1, 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). 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 the growth of χ^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 $\chi^l = \prod_k \delta\chi^k = \exp\left(\sum_k \log \delta\chi^l\right)$ 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 χ^l instead of the exponential evolution for feedforward nets. Equivalently, when rewriting Eq. (15):

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

the evolution of χ^l for resnets is the same as 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 χ^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 networks, batch normalization and skip connections.

Most studies until now have 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 assumption of $\mathbf{y}^l$ being 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 imply that for any given $\mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$: $\mathbb{E}_{\Theta^l}[\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}}] = \mathbb{E}_{\Theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^l)] = 0$, $\mathbb{E}_{\Theta^l}[(\mathbf{z}^l_{\boldsymbol{\alpha},\mathrm{c}})^2] = \mathbb{E}_{\Theta^l}[\nu_{2,\mathrm{c}}(\mathbf{z}^l)] = 1$, 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 $l \to \infty$.

In a similar vein: Duvenaud et al. (2014) found that the composition of Gaussian processes eventually lead to log-normal 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 log-normally distributed (with fat-tail) as $l \to \infty$.

Our approach also enabled 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 χ^l, this results in the convergence to a pathology with one-dimensional signal: $r_{\text{eff}}(\mathbf{x}^l) \to 1$ and with linear upper layers. The pathology of one-dimensional signal being invariant to constant rescalings, this reveals that the full hypothesis space with zero biases is pathological.

– 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 directions; on the other hand, ϕ. Again this pathology being invariant to constant rescalings, this reveals that the full hypothesis space with zero batch normalization shift is pathological.

– Finally in the case of resnets, χ^l only grows as a power-law. Equivalently, the evolution with depth of resnets is the logarithmic version of the evolution with depth of feedforward nets. The underlying phenomenon is the dilution of the residual path into the skip-connection path with ratio of signal variances decaying as $1/(l+1)$. This mechanism is responsible for breaking the circle of depth multiplicativity which causes pathologies for feedforward nets.

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. URL https://books.google.de/books?id=z39jQgAACAAJ.

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.

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.

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. 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

All four experiments of Fig. 2, 3, 4, 5 were made on CIFAR-10 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 $\big(\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)\big)$ and $\big(\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)\big)$. The distributions of $\big(\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)\big)$ and $\big(\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)\big)$ shown in Fig. 2 were estimated using $10,000$ such realizations. The limited width, slightly smaller than standard values, had two purposes: (i) outlining the diffusion process, stronger for smaller width; (ii) limiting computation time in order to gather more realizations.

In Fig. 3, 4, 5, we increased the width to a realistic value of $N_l = 512$. 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 1σ 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 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, a choice of M in the range of standard batch sizes further had the benefit 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 the simple case of M' batch-normalized random points $(\mathbf{z}_0, \dots, \mathbf{z}_{M'})$. In our case, M' must be seen as proportional to M but as $M' > M$ since the data distribution depends on the input $\mathbf{x}$ *and* the spatial position $\boldsymbol{\alpha}$ with M corresponding solely to the number of inputs $\mathbf{x}$. 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', meaning that 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 $\mu_1(|\mathbf{z}^l|)$ gets even smaller; (iii) $\exp(\overline{m}_{\text{BN}}[\chi^l])$ and $\delta_{\text{BN}}\chi^l$ get larger (iv) $|\exp(\overline{m}_\phi[\chi^l]) - 1|$ and $|\delta_\phi \chi^l - 1|$ get smaller.

– The dynamics of $|\exp(\overline{m}_\phi[\chi^l]) - 1|$ at very low depth in Fig. 4, 5 is explained by the fact that input images from CIFAR-10 have of number of channels equal to 3, which is much smaller than the width $N_l = 384$. The signal is therefore ill-conditioned at very low depth and quickly gets better conditioned, implying that $|\exp(\overline{m}_\phi[\chi^l]) - 1|$ is non-negligible at very low depth and quickly gets vanishing. This dynamics is very 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 change the boundary conditions from periodic to reflective and to zero-padding. The evolution with reflective conditions was always equivalent to the evolution with periodic 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 since padded zeros injected new signal directions in receptive fields at the borders of convolutions; (ii) the evolution of batch-normalized feedforward nets and batch-normalized resnets were equivalent.

– We tested to change the dataset from CIFAR-10 to MNIST. In the case of MNIST, the random initial convolution of stride 2 reduced the spatial dimension from 28 to $n = 14$ and increased the width from 1 to N_0. The evolution was equivalent except that the signal was slightly more fat-tailed at very low depth due to the original images being more fat-tailed in MNIST than in CIFAR-10.

– Finally we tested to change the fuzz parameter ϵ of batch normalization. The experiments of Fig. 2, 3, 4, 5 used the standard value $\epsilon = 0.001$ but we found indistinguishable evolutions when using the value $\epsilon = 0$.

B. Complementary definitions and notations

In this section, we use again $\mathbf{v}^l$ as placeholder for a generic tensor at any step 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 of $\mathbf{v}^{l-1}$ having 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 the input $\mathbf{v}^{l-1}$ to the tensor $\mathrm{RF}(\mathbf{v}^{l-1}) \in \mathbb{R}^{n\times\cdots\times n\times K_l^d N_{l-1}}$, defined such that $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},:}$ is 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 the input $\mathbf{v}^{l-1}$. Strictly speaking, RF depend on l, but this is implied by the argument so we write RF for simplicity.

Receptive field vectors. The *receptive field vector* 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, the notation $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}, \mathbf{v}^{l-1}$ is overloaded in the expectation. Again ρ and $\hat{\rho}$ are strictly speaking dependent 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

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

where $\boldsymbol{W}^l \in \mathbb{R}^{N_l \times R_l}$ is the suitably reshaped matricial form of $\boldsymbol{\omega}^l$. To lighten notation, we write $\mathbf{y}^l = \boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1}) + \boldsymbol{\beta}^l$ 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. In the case of 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, tensor moments have the *same distribution with respect to θ^l for both propagations*. Furthermore $\forall \boldsymbol{\alpha}, \mathrm{c}$: $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} + \bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = \left|\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}\right|$ and $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}} \bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}} = 0$, implying $\forall \boldsymbol{\alpha}, \mathrm{c}$: $(\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 + (\bar{\mathbf{x}}^l_{\boldsymbol{\alpha},\mathrm{c}})^2 = (\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}})^2$. Thus we have

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

Now we 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) is obtained using $\mathrm{d}\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}} = -\mathrm{d}\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}} = -\bar{\mathbf{y}}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and 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

$$\begin{aligned} \bar{\mathbf{y}}^l &= -\boldsymbol{W}^l \mathrm{RF}(\mathbf{x}^{l-1}) - \boldsymbol{\beta}^l, & \mathrm{d}\bar{\mathbf{y}}^l &= -\boldsymbol{W}^l \mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1}), && (20) \\ \bar{\mathbf{z}}^l &= \mathrm{BN}(\bar{\mathbf{y}}^l), & \mathrm{d}\bar{\mathbf{z}}^l &= \mathrm{BN}'(\bar{\mathbf{y}}^l) \odot \mathrm{d}\bar{\mathbf{y}}^l, && (21) \\ \bar{\mathbf{x}}^l &= \phi(\bar{\mathbf{z}}^l), & \mathrm{d}\bar{\mathbf{x}}^l &= \phi'(\bar{\mathbf{z}}^l) \odot \mathrm{d}\bar{\mathbf{z}}^l. && (22) \end{aligned}$$

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

$$\bar{\mathbf{z}}^l = -\mathbf{z}^l, \qquad \mathrm{d}\bar{\mathbf{z}}^l = -\mathrm{d}\mathbf{z}^l. \tag{23}$$

The same analysis as before gives

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

$$\forall \mathrm{c}: \ \mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{x}^l) + \mu_{2,\mathrm{c}}(\mathrm{d}\bar{\mathbf{x}}^l) = \mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{z}^l). \tag{25}$$

B.4. Gramian and Covariance

We adopt the standard definitions of the Gramian matrix and Covariance matrix of random vectors with respect to $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}$.

The *Gramian matrices* $\boldsymbol{G}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}$ 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

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

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

First we prove that RF is statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^{l-1}$ when convolutions have periodic boundary conditions and the global spatial extent n is constant and afterwards we provide a possible relaxation of these assumptions. When it is not constant, the global spatial extent is denoted n_l.

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* 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 $i_c \in \mathcal{I}_c^l$, and consider the tensors $\mathbf{v}^{l-1}_{:,c}$, $\mathrm{RF}(\mathbf{v}^{l-1})_{:,i_c} \in \mathbb{R}^{n\times\cdots\times n}$. The index i_c corresponds to a given convolution kernel position $\boldsymbol{\kappa} \in \{1,\dots,K_l\}^d$. Under periodic boundary conditions, this fixed kernel position implies that each position $\boldsymbol{\alpha}$ in $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_c}$ originates from a different position $\boldsymbol{\alpha}'$ in the tensor $\mathbf{v}^{l-1}_{\boldsymbol{\alpha}',c}$. 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 $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_c} = \mathbf{v}^{l-1}_{f(\boldsymbol{\alpha}),c} \sim_{\boldsymbol{\alpha}} \mathbf{v}^{l-1}_{\boldsymbol{\alpha},c}$ when $\mathbf{v}^{l-1}$ is deterministic and $\boldsymbol{\alpha}$ is random. In turn, this implies that $\mathrm{RF}(\mathbf{v}^{l-1})_{\boldsymbol{\alpha},i_c} \sim_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}} \mathbf{v}^{l-1}_{\boldsymbol{\alpha},c}$, 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* RF *is statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^{l-1}$. It follows in particular that* RF *is always statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^{l-1}$ in the fully-connected case $n_l = 1$.*

Proof. This follows immediately from Lemma 1. □

Corollary 3. *If convolutions have periodic boundary conditions and the global spatial extent n is constant, then* RF *is statistics-preserving with respect to $\mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^{l-1}$.*

Thus for any channel c *and $i_c \in \mathcal{I}_c^l$, we have: $\rho(\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_c} \sim_{\mathbf{x},\boldsymbol{\alpha}} \varphi(\mathbf{x}^{l-1},\boldsymbol{\alpha})_c$ and $\rho(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})_{i_c} \sim_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}} \varphi(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})_c$. Since the cardinality $|\mathcal{I}_c^l| = K_l^d$ is the same for all channels* c*, this implies 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 $\rho(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha}) = \varphi(\mathrm{d}\mathbf{x}^{l-1},\boldsymbol{\alpha})$ 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,* 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 $\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = \mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ and $\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:} = \mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},:}$ for most $\boldsymbol{\alpha}$. This implies for any index i_c that $P_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_c}\big] \simeq P_{\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_c}\big]$ and $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_c}\big] \simeq P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\big[\mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_c}\big]$.

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 that for any channel c and index $i_{\mathrm{c}} \in \mathcal{I}_{\mathrm{c}}^{l}$, we have $P_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{RF}^{(\mathrm{p})}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\right] = P_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$ and $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{RF}^{(\mathrm{p})}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\right] = P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{d}\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$. We then deduce that $P_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{RF}(\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\right] \simeq P_{\mathbf{x},\boldsymbol{\alpha}}\left[\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$ and $P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{RF}(\mathrm{d}\mathbf{x}^{l-1})_{\boldsymbol{\alpha},i_{\mathrm{c}}}\right] \simeq P_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}\left[\mathrm{d}\mathbf{x}^{l-1}_{\boldsymbol{\alpha},\mathrm{c}}\right]$, 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. Assumption that Φ_l is differentiable a.s. with respect to $\mathbf{x}$

Several results of Section 3 rely on the assumption that Φ_l is differentiable a.s. with respect to $\mathbf{x}$.

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

Secondly the *sensitivity equivalence* and the *Jabobian equivalence* detailed respectively in Sections C.2 and C.3 rely 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 again that moments with respect to $\mathbf{x}, \mathrm{d}\mathbf{x}, \boldsymbol{\alpha}$ are left unchanged when ignoring the probability-zero event under which $\Phi_l(\mathbf{x})$ is non-differentiable with respect to $\mathbf{x}$.

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

For given $\mathbf{x}$ such that $\forall \boldsymbol{\alpha}$: $\mathbf{x}_{\boldsymbol{\alpha},:} \neq 0$, it is easy to see that $\Phi_l(\mathbf{x})$ is *non-differentiable* implies that $\exists k \leq l$, $\exists \boldsymbol{\alpha}, \mathrm{c}$: $\rho(\mathbf{x}^{k-1}, \boldsymbol{\alpha}) \neq 0$ and $\mathbf{x}^k_{\boldsymbol{\alpha},\mathrm{c}} = 0$. Under standard initialization, this corresponds to a zero-probability event with respect to Θ^l. Denoting $D_{\mathbf{x}}$ the event such that $\Phi_l(\mathbf{x})$ is non-differentiable, we then have $\mathbb{P}_{\Theta^l|\mathbf{x}}[D_{\mathbf{x}}] = 0$. Now considering $\mathbf{x}$ again as random, using Fubini's Theorem and making the assumption that $\mathbf{x}_{\boldsymbol{\alpha},:} \neq 0$ a.s. with respect to $\mathbf{x}, \boldsymbol{\alpha}$ (which is the case e.g. if $\mathbf{x}_{\boldsymbol{\alpha},:}$ has well-defined probability density function):

$$\mathbb{E}_{\Theta^l}\left[\mathbb{P}_{\mathbf{x}|\Theta^l}[D_{\mathbf{x}}]\right] = \mathbb{E}_{\Theta^l}\mathbb{E}_{\mathbf{x}|\Theta^l}[\mathbf{1}_{D_{\mathbf{x}}}] = \mathbb{E}_{\mathbf{x}}\mathbb{E}_{\Theta^l|\mathbf{x}}[\mathbf{1}_{D_{\mathbf{x}}}] = \mathbb{E}_{\mathbf{x}}\left[\mathbb{P}_{\Theta^l|\mathbf{x}}[D_{\mathbf{x}}]\right] = 0. \tag{26}$$

By contradiction, if there would be non-zero probability with respect to Θ^l that $\mathbb{P}_{\mathbf{x}|\Theta^l}[D_{\mathbf{x}}] > 0$, then Eq. (26) would not hold. Therefore with probability 1 with respect to Θ^l: $\mathbb{P}_{\mathbf{x}|\Theta^l}[D_{\mathbf{x}}] = 0$, meaning that with probability 1 with respect to Θ^l, $\Phi_l(\mathbf{x})$ is differentiable a.s. with respect to $\mathbf{x}$.

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

Proof. Due to the definition of $\mathrm{d}\mathbf{x}^l$ as a small corruption to $\mathbf{x}^l$, $\mathrm{d}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ can be written as a function of $\nabla_{\mathbf{x}}\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and $\mathrm{d}\mathbf{x} = \mathrm{d}\mathbf{x}^0$:

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

with $\langle\,,\rangle$ the standard dot product in $\mathbb{R}^{n^d N_0}$. Then due to the property $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i \mathrm{d}\mathbf{x}_j] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x})\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x}^0)\delta_{ij}$:

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

□

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

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

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

C.3. Equivalence between χ^l and previous definitions

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

$$\left(\frac{\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right]}{N_l}\frac{\mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^0_{\mathrm{c}}]\right]}{\mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\right]}\right)^{\frac{1}{2}}. \tag{27}$$

Let us prove the equivalence between the definitions of Eq. (3) and Eq. (27). In the fully-connected case $n = 1$, the spatial position $\boldsymbol{\alpha}$ can be omitted and tensors and feature map vectors coincide: $\mathbf{x}^l = \varphi(\mathbf{x}^l)$ and $\mathrm{d}\mathbf{x}^l = \varphi(\mathrm{d}\mathbf{x}^l) = \hat{\varphi}(\mathrm{d}\mathbf{x}^l)$. When $\mathbf{x}$ is fixed, the input-output Jacobian $\boldsymbol{J}^l = \frac{\partial \mathbf{x}^l}{\partial \mathbf{x}^0} \in \mathbb{R}^{N_l \times N_0}$ directly summarizes the propagation of the noise: $\mathrm{d}\mathbf{x}^l = \boldsymbol{J}^l \mathrm{d}\mathbf{x} = \boldsymbol{J}^l \mathrm{d}\mathbf{x}^0$. Then due to the white noise property: $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_{\mathrm{c}}\mathrm{d}\mathbf{x}_{\mathrm{c}'}] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{\mathrm{c}\mathrm{c}'} = \mu_2(\mathrm{d}\mathbf{x})\delta_{\mathrm{c}\mathrm{c}'} = \mu_2(\mathrm{d}\mathbf{x}^0)\delta_{\mathrm{c}\mathrm{c}'}$, we deduce that

$$\begin{aligned}
\mathbb{E}_{\mathrm{d}\mathbf{x}}\Big[\sum\nolimits_{\mathrm{c}}(\mathrm{d}\mathbf{x}^l_{\mathrm{c}})^2\Big] &= \mathbb{E}_{\mathrm{d}\mathbf{x}}\Big[\,\mathrm{Tr}\,\mathrm{d}\mathbf{x}^l(\mathrm{d}\mathbf{x}^l)^T\Big] = \mathrm{Tr}\,\boldsymbol{J}^l\mathbb{E}_{\mathrm{d}\mathbf{x}}\Big[\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}^T\Big](\boldsymbol{J}^l)^T = \mu_2(\mathrm{d}\mathbf{x})\,\mathrm{Tr}\,\boldsymbol{J}^l(\boldsymbol{J}^l)^T = \mu_2(\mathrm{d}\mathbf{x})||\boldsymbol{J}^l||_F^2,\\
\frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x}^0)} &= \frac{\mu_2(\mathrm{d}\mathbf{x}^l)}{\mu_2(\mathrm{d}\mathbf{x})} = \frac{1}{\mu_2(\mathrm{d}\mathbf{x})}\mathbb{E}_{\mathbf{x},\mathrm{d}\mathbf{x},\mathrm{c}}\left[(\mathrm{d}\mathbf{x}^l_{\mathrm{c}})^2\right] = \frac{1}{\mu_2(\mathrm{d}\mathbf{x})}\mathbb{E}_{\mathbf{x}}\mathbb{E}_{\mathrm{d}\mathbf{x}}\left[\frac{1}{N_l}\sum\nolimits_{\mathrm{c}}(\mathrm{d}\mathbf{x}^l_{\mathrm{c}})^2\right] = \frac{1}{N_l}\mathbb{E}_{\mathbf{x}}\left[||\boldsymbol{J}^l||_F^2\right].
\end{aligned} \tag{28}$$

On the other hand, $\mu_2(\mathbf{x}^l)$ is defined as

$$\mu_2(\mathbf{x}^l) = \mathbb{E}_{\mathbf{x},\mathrm{c}}\left[\hat{\varphi}(\mathbf{x}^l)^2_{\mathrm{c}}\right] = \mathbb{E}_{\mathrm{c}}\left[\mathrm{Var}_{\mathbf{x}}[\mathbf{x}^l_{\mathrm{c}}]\right], \tag{29}$$

Combining Eq. (28) and Eq. (29), we get the equivalence between the two definitions:

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

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

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

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

The pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \xrightarrow{l\to\infty} 0$ 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 the multiplicatively centered increments as $\underline{\delta}\nu_2(\mathbf{x}^k) \equiv \delta\nu_2(\mathbf{x}^k)/\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]$, the term $\underline{m}[\nu_2(\mathbf{x}^k)]$ can be expressed as

$$\begin{aligned}\underline{m}[\nu_2(\mathbf{x}^k)] &= \mathbb{E}_{\theta^k}\Big[\log\Big(\underline{\delta}\nu_2(\mathbf{x}^k)\,\mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]\Big)\Big] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)]\\ &= \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)] + \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] - \log \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)], \qquad (30)\\ &= \mathbb{E}_{\theta^k}[\log \underline{\delta}\nu_2(\mathbf{x}^k)],\end{aligned}$$

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. (30). The term $\underline{s}[\nu_2(\mathbf{x}^k)]$ can be expressed as

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

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. (31).

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 $\mathbb{P}[X_k \le a] \xrightarrow{k\to\infty} \mathbb{P}[X \le a]$ for every continuity point a of the function $x \mapsto \mathbb{P}[X \le x]$. We then write $X_k \Rightarrow X$.

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

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

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

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

Lemma 7 (Theorem 25.7 in Billingsley (1995)). *Consider a real-valued function h, continuous everywhere apart from a finite set of discontinuity points $D_h = \{x_1, \dots, x_p\}$. In this case, 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. *Consider a sequence $(X_k)_k$ of random variables which depend on Θ^k and denote*

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

Further suppose that there exist constants m_{min}, m_{max}, v_{min}, v_{max} such that

$$\forall k \leq l:\ m_{\text{min}} \leq Y_k \leq m_{\text{max}}, \qquad \forall k \leq l:\ v_{\text{min}} \leq \text{Var}_{\theta^k}[Z_k] \leq v_{\text{max}}.$$

Then it follows that

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

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

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

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

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

$$\begin{aligned} \mathbb{E}_{\theta^k}[Z_k] &= \mathbb{E}_{\theta^k}[X_k] - \mathbb{E}_{\theta^k}[X_k] = 0, \\ \mathbb{E}_{\Theta^k}[Z_k] &= \mathbb{E}_{\Theta^{k-1}}[\mathbb{E}_{\theta^k}[Z_k]] = 0. \end{aligned} \tag{32}$$

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

$$\begin{aligned} \mathbb{E}_{\Theta^{k'}}[Z_k Z_{k'}] &= \mathbb{E}_{\Theta^{k'-1}} \mathbb{E}_{\theta^{k'}}[Z_k Z_{k'}] \\ &= \mathbb{E}_{\Theta^{k'-1}}\left[Z_k\, \mathbb{E}_{\theta^{k'}}[Z_{k'}]\right] = 0, \end{aligned} \tag{33}$$

where Eq. (33) follows from Eq. (32). □

Proof of (ii). First we note that

$$\text{Var}_{\Theta^k}\left[Z_k\right] = \mathbb{E}_{\Theta^k}\left[Z_k^2\right] = \mathbb{E}_{\Theta^{k-1}} \mathbb{E}_{\theta^k}\left[Z_k^2\right] = \mathbb{E}_{\Theta^{k-1}} \text{Var}_{\theta^k}\left[Z_k\right].$$

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

$$v_{\text{min}} \leq \text{Var}_{\Theta^k}\left[Z_k\right] \leq v_{\text{max}}. \tag{34}$$

Now let us denote $M_l \equiv \sum_{k=1}^{l} Y_k$ and $S_l \equiv \sum_{k=1}^{l} Z_k$. Then we have

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

where we used (i) in Eq. (35). The hypothesis then implies $l m_{\min} \leq M_l \leq l m_{\max}$, while Eq. (34) and Eq. (35) together imply $l v_{\min} \leq \mathrm{Var}_{\Theta^l}[S_l] \leq l v_{\max}$. If we finally define $m_l \equiv M_l/l$ and $s_l \equiv S_l/\sqrt{l}$, the telescoping sum $\sum_{k=1}^{l} X_k = \sum_{k=1}^{l} Y_k + \sum_{k=1}^{l} Z_k$ can be written as required:

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

D.3. Proof of Theorem 1

Theorem 1 (moments of vanilla nets)**.** *There exist small constants $1 \gg m_{\min}, m_{\max}, v_{\min}, v_{\max} > 0$, 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*

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

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

D.3.1. Proof introduction

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

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

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

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

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

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) = u^l_{\mathrm{c}}\, R_l\, \nu_2(\mathbf{x}^{l-1}) \sum\nolimits_i \big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2 \hat{\lambda}_i, \tag{36}$$

$$\forall \mathrm{c}: \ \nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l) = R_l\, \nu_2(\mathbf{x}^{l-1}) \sum\nolimits_i \big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2 \hat{\lambda}_i. \tag{37}$$

Now combining Eq. (37) with the symmetry of the propagation: $\nu_{2,\mathrm{c}}(\mathbf{x}^l) \sim_{\theta^l} \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)$, and the assumption of standard initialization: $\boldsymbol{W}^l_{\mathrm{c},:} \sim_{\theta^l} \hat{\boldsymbol{W}}^l_{\mathrm{c},:} \sim_{\theta^l} \mathcal{N}(0, 2\,/\,R_l \boldsymbol{I})$, we get that under $\big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

$$\forall \mathrm{c}: \ 2\mathbb{E}_{\theta^l}\big[\nu_{2,\mathrm{c}}(\mathbf{x}^l)\big] = \mathbb{E}_{\theta^l}\big[\nu_{2,\mathrm{c}}(\mathbf{x}^l) + \nu_{2,\mathrm{c}}(\bar{\mathbf{x}}^l)\big] = \mathbb{E}_{\theta^l}\big[R_l\, \nu_2(\mathbf{x}^{l-1}) \sum\nolimits_i \big(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\big)^2 \hat{\lambda}_i\big]$$

$$= R_l\,\nu_2(\mathbf{x}^{l-1})\frac{2}{R_l}\sum\nolimits_i \hat{\lambda}_i = 2\nu_2(\mathbf{x}^{l-1}).$$

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

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

Let us next define

$$v_{\mathrm{c}}^l \equiv \begin{cases} 0 & \text{if } u_{\mathrm{c}}^l < \frac{1}{2} \\ 1 & \text{if } u_{\mathrm{c}}^l > \frac{1}{2} \\ b & \text{if } u_{\mathrm{c}}^l = \frac{1}{2} \end{cases}$$

with b independent of Θ^l and following a Bernouilli distribution with probability $1/2$: $b \sim \mathrm{Bernouilli}(1/2)$.

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

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

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

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

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 $\mathbb{P}_{\theta^l|B_l\cap\{\nu_2(\mathbf{x}^{l-1})\neq 0\}}\big[\nu_2(\mathbf{x}^l) \neq 0\big] = 1$, with B_l^{c} the complementary event of B_l.

Now since $(\nu_{2,\mathrm{c}}(\mathbf{y}^l))_{1\leq \mathrm{c}\leq N_l}$ and $(v_{\mathrm{c}}^l)_{1\leq \mathrm{c}\leq N_l}$ are independent, Eq. (36) implies that $\exists (w_i)_{1\leq i\leq R_l}$ such that under $B_l \cap \big\{\nu_2(\mathbf{x}^{l-1}) \neq 0\big\}$:

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

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

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

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

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

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

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

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.

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 $P_{\mathbf{x}}(\mathbf{x}) = P_{\mathbf{x}^0}(\mathbf{x}^0)$, $P_{\mathrm{d}\mathbf{x}}(\mathrm{d}\mathbf{x}) = P_{\mathrm{d}\mathbf{x}^0}(\mathrm{d}\mathbf{x}^0)$, and (ii) by Θ^{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 $(\mathbf{x}^{0,k})_{k\in\mathbb{N}}$ and $(\mathrm{d}\mathbf{x}^{0,k})_{k\in\mathbb{N}}$ corresponding to deterministic distributions $P_{\mathbf{x}^{0,k}}(\mathbf{x}^{0,k})$ and $P_{\mathrm{d}\mathbf{x}^{0,k}}(\mathrm{d}\mathbf{x}^{0,k})$ and sequences of deterministic $(\Theta^{l-1,k})_{k\in\mathbb{N}}$.
- The 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 deterministic parameters $\Theta^{l-1,k}$ such that $\nu_2(\mathbf{x}^{l-1,k}) > 0$ and $\mu_2(\mathrm{d}\mathbf{x}^{l-1,k}) > 0$.
- The 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 of $(\mathbf{x}^{l-1,k}, \mathrm{d}\mathbf{x}^{l-1,k})$ with random parameters $(\boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ and the 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 $\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k}) \equiv \frac{\mu_2(\mathrm{d}\mathbf{x}^{l,k})}{\mu_2(\mathrm{d}\mathbf{x}^{l-1,k})}$.
- The 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 then consider sequences such that $\mathbb{E}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})]$, $\mathrm{Var}_{\theta^l|A_{l,k}}[\log\delta\nu_2(\mathbf{x}^{l,k})]$ and $\mathbb{E}_{\theta^l|A'_{l,k}}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k})]$, $\mathrm{Var}_{\theta^l|A'_{l,k}}[\log\delta\mu_2(\mathrm{d}\mathbf{x}^{l,k})]$ 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 $B_{l,k} \cap A_{l-1,k}$:

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

with $\wedge$ the logical *and*, $\vee$ the logical *or*, and with $w_{\min,k}$, $w_{\max,k}$ defined as in Eq. (39) with respect to $\delta\nu_2(\mathbf{x}^{l,k})$. Then $\mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin]a,b]\big] = \mathbb{P}_{\theta^l|A_{l-1,k}\cap B_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \notin]a,b]\big]$ 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 $(\delta\nu_2(\mathbf{x}^{l,k}))_{k\in\mathbb{N}}$ of random variables, considered conditionally on $A_{l,k}$, is tight. By Lemma 8, this implies that there exists a sequence of strictly increasing indices $(i_k)_{k\in\mathbb{N}}$ and a random variable X such that $(\delta\nu_2(\mathbf{x}^{l,i_k}))_{k\in\mathbb{N}}$, considered conditionally on A_{l,i_k}, 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, we may thus rename without loss of generality the sequence $(\delta\nu_2(\mathbf{x}^{l,i_k}))_{k\in\mathbb{N}}$ 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 took the limit as $k \to \infty$ and used the definition of weak convergence $\mathbb{P}_{\theta^l|A_{l,k}}\big[\delta\nu_2(\mathbf{x}^{l,k}) \leq a\big] \xrightarrow{k\to\infty} \mathbb{P}[X \leq a]$.

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, and thus 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 $a_p \xrightarrow{p\to\infty} 0$. By Eq. (46), we then obtain $\mathbb{P}\big[X = 0\big] \leq \mathbb{P}\big[X \leq a_p\big] \xrightarrow{p\to\infty} 0$, and thus $\mathbb{P}\big[X = 0\big] = 0$. Without loss of generality, we may assume $X > 0$ surely (if it not the case, simply replace X by a constant arbitrary value > 0 under the zero-probability event $\{X = 0\}$).

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

D.3.3. UNIFORM INTEGRABILITY

Since $x \mapsto \mathbf{1}_{\{x \geq M\}} x$ is non-decreasing, Eq. (39) implies that

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

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

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

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

$$\begin{aligned}
\log w_{\text{min},k} &\leq \log \delta\nu_2(\mathbf{x}^{l,k}) \leq \log w_{\text{max},k}, \\
\left| \log \delta\nu_2(\mathbf{x}^{l,k}) \right| &\leq \max \left(\left| \log w_{\text{min},k} \right|, \left| \log w_{\text{max},k} \right| \right), \\
\mathbf{1}_{\{|\log \delta\nu_2(\mathbf{x}^{l,k})| \geq M\}} \left| \log \delta\nu_2(\mathbf{x}^{l,k}) \right| &\leq \max \left(\mathbf{1}_{\{|\log w_{\text{min},k}| \geq M\}} \left| \log w_{\text{min},k} \right|, \mathbf{1}_{\{|\log w_{\text{max},k}| \geq M\}} \left| \log w_{\text{max},k} \right| \right), \\
\mathbf{1}_{\{|\log \delta\nu_2(\mathbf{x}^{l,k})| \geq M\}} \left| \log \delta\nu_2(\mathbf{x}^{l,k}) \right| &\leq \mathbf{1}_{\{|\log w_{\text{min},k}| \geq M\}} \left| \log w_{\text{min},k} \right| + \mathbf{1}_{\{|\log w_{\text{max},k}| \geq M\}} \left| \log w_{\text{max},k} \right|.
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \left(\log \delta\nu_2(\mathbf{x}^{l,k}) \right)^2 &\leq \max \left(\mathbf{1}_{\{(\log w_{\text{min},k})^2 \geq M\}} \left(\log w_{\text{min},k} \right)^2, \mathbf{1}_{\{(\log w_{\text{max},k})^2 \geq M\}} \left(\log w_{\text{max},k} \right)^2 \right), \\
\mathbf{1}_{\{(\log \delta\nu_2(\mathbf{x}^{l,k}))^2 \geq M\}} \left(\log \delta\nu_2(\mathbf{x}^{l,k}) \right)^2 &\leq \mathbf{1}_{\{(\log w_{\text{min},k})^2 \geq M\}} \left(\log w_{\text{min},k} \right)^2 + \mathbf{1}_{\{(\log w_{\text{max},k})^2 \geq M\}} \left(\log w_{\text{max},k} \right)^2.
\end{aligned}$$

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

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

It follows that

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

This means that $(\log \delta\nu_2(\mathbf{x}^{l,k}))_{k \in \mathbb{N}}$, $((\log \delta\nu_2(\mathbf{x}^{l,k}))^2)_{k \in \mathbb{N}}$ are uniformly integrable conditionally on $A_{l,k}$, and by Lemma 9:

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

D.3.4. BOUNDING MOMENTS OF $\delta\nu_2(\mathbf{x}^{l,k})$

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

$$\begin{aligned}\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})}{\nu_2(\mathbf{x}^{l-1,k})}\right] &\leq \mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})^2}{\nu_2(\mathbf{x}^{l-1,k})^2}\right] \\ &\leq \mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\nu_{2,\mathrm{c}}(\mathbf{y}^{l,k})^2}{\nu_2(\mathbf{x}^{l-1,k})^2}\right] \\ &\leq R_l^2\mathbb{E}_{\theta^l|A_{l-1,k}}\Big[\sum\nolimits_{i\neq i'}\left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^2\left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i'}\right)^2\hat{\lambda}_i\hat{\lambda}_{i'}+\sum\nolimits_i\left(\hat{\boldsymbol{W}}^l_{\mathrm{c},i}\right)^4\hat{\lambda}_i^2\Big] \\ &\leq R_l^2\sum\nolimits_{i,i'}\left(\frac{2}{R_l}\right)\left(\frac{2}{R_l}\right)3\hat{\lambda}_i\hat{\lambda}_{i'}=12.\end{aligned} \tag{47}$$

Since the different channels $\nu_{2,\mathrm{c}}(\mathbf{x}^{l,k})_\mathrm{c}$ are mutually independent, we get

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

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

$$\begin{aligned}&\left|\mathbb{E}_{\theta^l|A_{l,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]-1\right| \\ &\quad=\left|\mathbb{E}_{\theta^l|A_{l,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]-\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]\right| \\ &\quad=\left|\left(\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]^{-1}-1\right)\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\mathbf{1}_{A_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\right]-\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\mathbf{1}_{A^\mathrm{c}_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\right]\right| \\ &\quad\leq\frac{\mathbb{P}_{\theta^l|A_{l-1,k}}[A^\mathrm{c}_{l,k}]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}\left|\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\mathbf{1}_{A_{l,k}}\delta\nu_2(\mathbf{x}^{l,k})\right]\right|+\mathbb{P}_{\theta^l|A_{l-1,k}}\left[A^\mathrm{c}_{l,k}\right]^{\frac{1}{2}}\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})^2\right]^{\frac{1}{2}} &(48)\\ &\quad\leq\left(\frac{\mathbb{P}_{\theta^l|A_{l-1,k}}[A^\mathrm{c}_{l,k}]}{\mathbb{P}_{\theta^l|A_{l-1,k}}[A_{l,k}]}+\mathbb{P}_{\theta^l|A_{l-1,k}}\left[A^\mathrm{c}_{l,k}\right]^{\frac{1}{2}}\right)\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})^2\right]^{\frac{1}{2}} &(49)\\ &\quad\leq\left(\frac{2^{-N_l}}{1-2^{-N_l}}+2^{-\frac{N_l}{2}}\right)\left(1+\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]\right)^{\frac{1}{2}} \\ &\quad\leq\epsilon_l\left(1+\frac{12}{N_l}\right)^{\frac{1}{2}}\leq 2\epsilon_l, &(50)\end{aligned}$$

where we applied Cauchy-Schwarz inequality in Eq. (48) and Eq. (49), we defined $\epsilon_l\equiv\frac{2^{-N_l}}{1-2^{-N_l}}+2^{-\frac{N_l}{2}}$ and we used $\left(1+\frac{12}{N_l}\right)^{\frac{1}{2}}\leq 2$ under the large width assumption.

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

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

$$\leq \frac{1}{1-2^{-N_l}}\left(\frac{12}{N_l}+2^{-N_l}\right)+2\epsilon_l\left(\frac{1}{1-2^{-N_l}}+1\right)\leq \frac{24}{N_l}, \tag{51}$$

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

Finally let us bound $\mathrm{Var}_{\theta^l|A_{l,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]$ from below. Under the conditionality $||\boldsymbol{W}^l||_F^2>0$, we have: (i) $||\boldsymbol{W}^l||_F$ is independent of $B_{l,k}$; (ii) $||\boldsymbol{W}^l||_F$ is independent of $\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}$ by spherical symmetry. In the remaining of the calculation to bound $\mathrm{Var}_{\theta^l|A_{l,k}}\left[\delta\nu_2(\mathbf{x}^{l,k})\right]$, the conditionality on the event $\left\{||\boldsymbol{W}^l||_F^2>0\right\}$ is assumed but is omitted for simplicity of notation. This conditionality has no effect on expectations and probabilities since the event $\left\{||\boldsymbol{W}^l||_F^2>0\right\}$ has probability one.

We have

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

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

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

It remains to bound the terms $\mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]^2$ and $\mathrm{Var}_{\theta^l|A_{l-1,k}}\left[||\boldsymbol{W}^l||_F^2\right]$. A computation similar to Eq. (49) gives

$$\left|\mathbb{E}_{\theta^l|A_{l,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]-\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]\right|\leq\epsilon_l\,\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\right]^{\frac{1}{2}}. \tag{53}$$

The term $\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})^2}{||\boldsymbol{W}^l||_F^4}\right]^{\frac{1}{2}}$ of Eq. (53) can be bounded using Eq. (36):

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

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

As for the term $\mathbb{E}_{\theta^l|A_{l-1,k}}\left[\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}\right]$ of Eq. (53), we use the independence of $||\boldsymbol{W}^l||_F$ and $\frac{\delta\nu_2(\mathbf{x}^{l,k})}{||\boldsymbol{W}^l||_F^2}$ to get the identity:

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

We have $\epsilon_l = \frac{2^{-N_l}}{1-2^{-N_l}} + 2^{-\frac{N_l}{2}} \leq 2\times 2^{-N_l} + 2^{-\frac{N_l}{2}} \ll \frac{1}{2N_{l-1}n^d} \leq \frac{1}{2R_l}$ under the assumption of large width. Eq. (53) then gives

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

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

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

Finally combining Eq. (52), Eq. (54) and Eq. (55):

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

D.3.5. CONSEQUENCE FOR m_{min}, m_{max}, v_{min}, v_{max}

Using Eq. (50) and taking the limit $k\to\infty$:

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

Similarly, using Eq. (51) and Eq. (56) and taking the limit $k\to\infty$:

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

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: $\frac{1}{\sqrt{2N_lR_l}} \leq \mathrm{Var}\big[X\big]^{\frac{1}{2}} \leq \sqrt{\frac{24}{N_l}}$. This means that $\big|\mathbb{E}\big[X\big] - 1\big|$ is much smaller than the effect of the log-concavity:

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

In addition, X has small standard deviation around 1 since $\mathrm{Var}\big[X\big]^{\frac{1}{2}} \ll 1$ under the assumption of large width, and thus

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

Now if we alternately consider sequences $(\delta\nu_2(\mathbf{x}^{l,k}))_{k\in\mathbb{N}}$ 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 $\Theta^{l-1,k}$ such that

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

then we obtain

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

The final remaining dependency is the dependency in N_l and R_l. Given that $(N_l)_{l\in\mathbb{N}}$ is bounded, and given $R_l = K_l^d N_{l-1} \leq n^d N_{l-1}$, it follows that $(R_l)_{l\in\mathbb{N}}$ is also bounded. If we denote $N_{\min} \equiv \min_l N_l$, $N_{\max} \equiv \max_l N_l$ $R_{\min} \equiv \min_l R_l$, $R_{\max} \equiv \max_l R_l$, as well as $\mathcal{I}_N \equiv \{N_{\min}, \ldots, N_{\max}\}$ and $\mathcal{I}_R \equiv \{R_{\min}, \ldots, R_{\max}\}$, we finally get

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

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

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

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

$$m_{\min} \leq \mathbb{E}_{\theta^l|A_l}[-\log \delta\nu_2(\mathbf{x}^l)] \leq m_{\max}, \qquad v_{\min} \leq \mathrm{Var}_{\theta^l|A_l}[\log \delta\nu_2(\mathbf{x}^l)] \leq v_{\max}, \tag{57}$$
$$m_{\min} \leq \mathbb{E}_{\theta^l|A'_l}[-\log \delta\mu_2(\mathrm{d}\mathbf{x}^l)] \leq m_{\max}, \qquad v_{\min} \leq \mathrm{Var}_{\theta^l|A'_l}[\log \delta\mu_2(\mathrm{d}\mathbf{x}^l)] \leq v_{\max}. \tag{58}$$

D.3.6. PROOF CONCLUSION

Again we start by focusing on $\delta\nu_2(\mathbf{x}^l)$, and the reasoning will be easily extended to $\delta\mu_2(\mathrm{d}\mathbf{x}^l)$. From now on, we work under A_l, and we define

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

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

$$\begin{aligned} Y_k &= \mathbb{E}_{\theta^k|A_l}[\log \delta\nu_2(\mathbf{x}^k)] = \mathbb{E}_{\theta^k|A_k}[\log \delta\nu_2(\mathbf{x}^k)], \\ \mathrm{Var}_{\theta^k|A_l}[Z_k] &= \mathbb{E}_{\theta^k|A_l}\left[Z_k^2\right] = \mathbb{E}_{\theta^k|A_k}\left[\left(\log \delta\nu_2(\mathbf{x}^k) - \mathbb{E}_{\theta^k|A_k}[\log \delta\nu_2(\mathbf{x}^k)]\right)^2\right] = \mathrm{Var}_{\theta^k|A_k}[\log \delta\nu_2(\mathbf{x}^k)]. \end{aligned}$$

Using Eq. (57), it follows that $\forall k \leq l$ under A_l:

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

Applying Lemma 10 conditionally on A_l, we deduce that there exist random variables m_l, s_l such that under A_l:

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

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

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

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:

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

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 $m_{\min}$, $m_{\max}$, $v_{\min}$, $v_{\max}$ are obtained by considering extreme cases for u_{c}^l and $R_l \sum_i (\hat{\boldsymbol{W}}_{\mathrm{c},i}^l)^2 \hat{\lambda}_i$ in Eq. (36):

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

We numerically find $m_{\min} \simeq 9.7{\times}10^{-5}$ and $v_{\min} \simeq 2.0{\times}10^{-4}$ as minimum bounds and $m_{\max} \simeq 2.5{\times}10^{-2}$ and $v_{\max} \simeq 5.2{\times}10^{-2}$ as maximum bounds.

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

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

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

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

D.5. Relation to the terms $\overline{m}$, $\underline{m}$, $\underline{s}$ defined in Section 4

Here we relate Theorem 1 to the terms $\overline{m}$, $\underline{m}$, $\underline{s}$ defined in Section 4, under the conditionality A_k, A'_k. By Eq. (50), we have $|\mathbb{E}_{\theta^k|A_k}[\delta\nu_2(\mathbf{x}^k)] - 1| \ll 1$. This implies that under A_k:

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

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

$$|\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^k)]| = \left|\log \mathbb{E}_{\theta^k|A'_k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)]\right| \simeq \left|\mathbb{E}_{\theta^k|A'_k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)] - 1\right| \ll 1,$$

The terms $\overline{m}[\nu_2(\mathbf{x}^k)]$ and $\overline{m}[\mu_2(\mathrm{d}\mathbf{x}^k)]$ are thus vanishing and the evolution with depth of $\nu_2(\mathbf{x}^l), \mu_2(\mathrm{d}\mathbf{x}^l)$ is dominated by the terms $\underline{m}[\nu_2(\mathbf{x}^k)] < 0$, $\underline{m}[\mu_2(\mathrm{d}\mathbf{x}^k)] < 0$.

D.6. Proof of Theorem 2

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

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

Proof. The dominating term in the evolution of χ^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 $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$ and $(\lambda_1, \ldots, \lambda_{R_l})$ respectively the orthogonal eigenvectors and eigenvalues of $\boldsymbol{C}_{\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathbf{x}^{l-1}, \boldsymbol{\alpha})]$ and $\hat{\boldsymbol{W}}^l \equiv \boldsymbol{W}^l(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{R_l})$. Using these notations, we get

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

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

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

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

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

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

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) is obtained by symmetry of the propagation. We then get

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

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

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

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: $4\mathbb{E}_{\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})] \leq 2\mu_2(\mathbf{x}^{l-1})$, and thus

$$\mathbb{E}_{\mathrm{c},\theta^l}\left[\frac{\nu_{1,\mathrm{c}}(\mathbf{y}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{y}^{l,-})}{\mu_2(\mathbf{x}^{l-1})}\right] \leq \frac{1}{2},$$

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

D.7. If χ^l has exponential drift larger than diffusion and $\mu_2(\mathbf{x}^l)$, $\nu_2(\mathbf{x}^l)$ are lognormal, then $\mu_2(\mathbf{x}^l) \,/\, \nu_2(\mathbf{x}^l) \to 0$ a.s.

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

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

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

Now let us reason by contradiction and suppose $\exists E$ with $\mathbb{P}[E] > 0$ such that, under E: $X_l \xrightarrow{l\to\infty\!\!\!\!\!/} X$. Under E, $\exists \epsilon$ random variable and $\exists (k_l)_{l\in\mathbb{N}}$ random strictly increasing sequence with $\forall l$: $|X_{k_l} - X| > \epsilon$. This implies in turn $\exists E'$ 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. $\square$

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 χ^l has drift larger than diffusion has probability $\mathbb{P}[D] > 0$

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

Then, under D:

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

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

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

On the other hand, under standard initialization:

$$\mathbb{E}_{\Theta^l}[\nu_2(\mathbf{x}^l)] = \nu_2(\mathbf{x}^0) \prod_{k=1}^{l} \mathbb{E}_{\theta^k}[\delta\nu_2(\mathbf{x}^k)] = \nu_2(\mathbf{x}^0),$$
$$\mathbb{E}_{\Theta^l}[\mu_2(\mathrm{d}\mathbf{x}^l)] = \mu_2(\mathrm{d}\mathbf{x}^0) \prod_{k=1}^{l} \mathbb{E}_{\theta^k}[\delta\mu_2(\mathrm{d}\mathbf{x}^k)] = \mu_2(\mathrm{d}\mathbf{x}^0).$$

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

$\exists M_l, M'_l > 0$ constants such that

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

Now let us make more precise the conditionality on D. We may assume $\exists m > \frac{1}{2}\big(m_{\max} - m_{\min}\big)$ 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

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

which gives with logarithms that, under D:

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

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

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

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

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

Then for l large enough:

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

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 $\operatorname{erfc}(x) \leq \exp(-x^2)$ (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 $\exp\left(\overline{m}[\chi^l]\right) \to 1$ and $\tilde{\mathbf{x}}^l$ has bounded moments, then $\mathbf{x}^l$ converges to one-dimensional signal pathology

Proposition 13. *Again we adopt the notation for the unit-variance rescaled signal: $\tilde{\mathbf{x}}^l \equiv \mathbf{x}^l/\sqrt{\mu_2(\mathbf{x}^l)}$, 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 $\nu_2(\mathbf{x}^l)/\mu_2(\mathbf{x}^l) \not\xrightarrow{l\to\infty} \infty$ and thus $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \not\xrightarrow{l\to\infty} 0$, i.e. that $\mathbf{x}^l$ does not converge to zero-dimensional signal pathology,*
- *(ii) $\delta\chi^l \simeq \exp\left(\overline{m}[\chi^l]\right) \xrightarrow{l\to\infty} 1$.*

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

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

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

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

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

i.e. $||\boldsymbol{\nu}^l_\rho||_2^2 = R_l\big(\nu_2(\tilde{\mathbf{x}}^l) - 1\big)$. 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 $r_{\text{eff}}(\mathbf{x}^l) = r_{\text{eff}}(\tilde{\mathbf{x}}^l) \not\xrightarrow{l\to\infty} 1$, which implies $\exists \eta > 0$ and $\exists (k_l)_{l\in\mathbb{N}}$ strictly increasing sequence with $\forall l$: $r_{\text{eff}}(\tilde{\mathbf{x}}^{k_l}) \geq 1 + \eta$. In turn this implies $\exists \eta' > 0$ such that $\forall l$:

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

i.e. that $\varphi(\tilde{\mathbf{x}}^{k_l}, \boldsymbol{\alpha})$ necessarily 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 $\forall l$:

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

We 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 $\tilde{\boldsymbol{\nu}}^{k_l}_{\rho} \equiv \boldsymbol{\nu}^{k_l}_{\rho}/||\boldsymbol{\nu}^{k_l}_{\rho}||_2$. We 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^2_{\mathbf{v}}\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}^{k_l}_{\rho}\rangle^2 + (1-w^2_{\mathbf{v}})\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle^2 + 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}^{k_l}_{\rho}\rangle\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big] \\
&\quad\geq w^2_{\mathbf{v}}\eta' + 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}^{k_l}_{\rho}\rangle\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big] \\
&\quad\geq w^2_{\mathbf{v}}\eta' - 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{v}^{k_l}_{\rho}\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^2_{\mathbf{v}}\eta' - 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}\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^2_{\mathbf{v}}\eta' - 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\sum\nolimits_i \rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^2\Big]^{\frac{1}{2}}\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\sum\nolimits_i \rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha})_i^2\Big]^{\frac{1}{2}} \\
&\quad\geq w^2_{\mathbf{v}}\eta' - 2w_{\mathbf{v}}\sqrt{1-w^2_{\mathbf{v}}}R_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^2_{\mathbf{v}})\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\Big[\langle\rho(\tilde{\mathbf{x}}^{k_l},\boldsymbol{\alpha}),\mathbf{w}\rangle\Big]^2 \leq (1-w^2_{\mathbf{v}})\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\right]^2 \\
&\quad\leq (1-w^2_{\mathbf{v}})\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] \leq (1-w^2_{\mathbf{v}})R_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 $\forall\epsilon > 0$, $\exists p_\epsilon$ 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

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

where Eq. (72) and Eq. (73) are obtained by again applying Cauchy-Schwarz inequality, while Eq. (74) is obtained with $\forall i, \forall p$: $\mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\rho(\tilde{\mathbf{x}}^{k_l}, \boldsymbol{\alpha})_i^p\right] \leq \sum_{\mathrm{c}} \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\varphi(\tilde{\mathbf{x}}^{k_l}, \boldsymbol{\alpha})_{\mathrm{c}}^p\right] = N_l \nu_p(\tilde{\mathbf{x}}^{k_l})$. It then follows from Eq. (71) and the hypothesis that all moments are bounded $\nu_p(|\tilde{\mathbf{x}}^l|) = \mathcal{O}(1)$ that

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

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

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

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

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

Let us note that

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

We then get $\exists \eta''' > 0$, $\exists p > 0$ such that $\forall l$:

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

We finally get

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

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

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

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

D.9. If $\exp\left(\overline{m}[\chi^l]\right) \to 1$, then each new layer l becomes arbitrary well approximated by a linear mapping

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

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

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

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

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

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

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

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

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

We have shown that $\mathbf{x}^l/\mu_2(\mathbf{x}^{l-1})$ is arbitrary 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 arbitrary well by a linear function of $\mathbf{x}^{l-1}$ when normalizing with respect to $\mathbf{x}^l$.

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

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

Using Eq. (65), we then get

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

Similarly to the proof of Theorem 1, we define

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

with $b \sim \text{Bernouilli}(1/2)$, independently of Θ^l. We further define $C_l \equiv \left\{\exists \text{c} : \ w_{\text{c}}^l = 1\right\}$ such that $\mathbb{P}_{\theta^l}[C_l] = 1 - 2^{-N_l}$. Since C_l is independent from $\mu_{2,\text{c}}(\mathbf{x}^l) + \mu_{2,\text{c}}(\bar{\mathbf{x}}^l)$ for all channels c, it follows from Eq. (78) that $\forall \eta > 0, \exists m > 0$ such that for l large enough:

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

It follows that $\exists m' > 0$ such that for l large enough:

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

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

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

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 ϵ by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrary close to $(1-\eta)\mathbb{P}_{\theta^l}[C_l] = (1-\eta)\left(1 - 2^{-N_l}\right)$. Thus $\tilde{\mathbf{x}}^l$ can be approximated abitrary well by a linear function of $\mathbf{x}^{l-1}$ with probability arbitrary 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 for large width $N_l \gg 1$.

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 χ^l can be decomposed as*

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

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

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

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

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

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

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

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

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

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

Term $\exp\left(\overline{m}_{\phi}\left[\chi^l\right]\right)$**.** We consider the symmetric propagation for batch-normalized feedforward nets, introduced in Section B. From Eq. (25), we deduce that

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}4\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-}) &\leq \mu_{2,\mathrm{c}}(\mathbf{z}^l) = 1,\\ 2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})] &\leq \frac{1}{2},\\ 1 \leq \exp\left(\overline{m}_\phi[\chi^l]\right) = \left(1 - 2\mathbb{E}_{\mathrm{c},\theta^l}[\nu_{1,\mathrm{c}}(\mathbf{z}^{l,+})\nu_{1,\mathrm{c}}(\mathbf{z}^{l,-})]\right)^{-\frac{1}{2}} &\leq \sqrt{2}.\end{aligned}$$

□

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

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

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

$$\begin{aligned}&= \sum\nolimits_{i,j} \boldsymbol{W}^1_{\mathrm{c},i}\boldsymbol{W}^1_{\mathrm{c},j}\mathbb{E}_{\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_i\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_j]\\ &= \mu_2(\mathrm{d}\mathbf{x}^0)\sum\nolimits_i \left(\boldsymbol{W}^1_{\mathrm{c},i}\right)^2 = \mu_2(\mathrm{d}\mathbf{x}^0)\,||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2.\end{aligned} \tag{87}$$

where Eq. (86) follows from $\mathrm{d}\mathbf{y}^1$ being centered and Eq. (87) follows from the white noise property $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\mathrm{d}\mathbf{x}_i\mathrm{d}\mathbf{x}_j] = \sigma^2_{\mathrm{d}\mathbf{x}}\delta_{ij} = \mu_2(\mathrm{d}\mathbf{x}^0)\,\delta_{ij}$, which implies for any $\boldsymbol{\alpha}$ that $\mathbb{E}_{\mathrm{d}\mathbf{x}}[\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_i\rho(\mathrm{d}\mathbf{x},\boldsymbol{\alpha})_j] = \mu_2(\mathrm{d}\mathbf{x}^0)\,\delta_{ij}$ under periodic boundary conditions.

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

$$\begin{aligned}\mu_{2,\mathrm{c}}(\mathbf{y}^1) = \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\hat{\varphi}(\mathbf{y}^1,\boldsymbol{\alpha})_{\mathrm{c}}^2\right] &= \mathbb{E}_{\mathbf{x},\boldsymbol{\alpha}}\left[\left(\boldsymbol{W}^1_{\mathrm{c},:}\hat{\rho}(\mathbf{x},\boldsymbol{\alpha})\right)^2\right] = \sum\nolimits_i \left(\hat{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i\\ &= ||\boldsymbol{W}^1_{\mathrm{c},:}||_2^2\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i = \frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^1)}{\mu_2(\mathrm{d}\mathbf{x}^0)}\sum\nolimits_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2\lambda_i,\end{aligned} \tag{88}$$

where we defined $\tilde{\boldsymbol{W}}^1$ such that $\forall \mathrm{c}$: $\tilde{\boldsymbol{W}}^1_{\mathrm{c},:} = \hat{\boldsymbol{W}}^1_{\mathrm{c},:} \,/\, ||\hat{\boldsymbol{W}}^1_{\mathrm{c},:}||$ 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 sphere of $\mathbb{R}^{R_1}$. In turn, this implies

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

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

$$\begin{aligned}
\exp\left(\overline{m}_{\mathrm{BN}}\left[\chi^1\right]\right) &= \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\mu_2(\mathbf{x}^0)}\right)^{-\frac{1}{2}} \mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{\mu_{2,\mathrm{c}}(\mathrm{d}\mathbf{y}^1)}{\mu_{2,\mathrm{c}}(\mathbf{y}^1)}\right]^{\frac{1}{2}} \\
&= \left(\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\frac{1}{R_1}\sum_i \lambda_i}\right)^{-\frac{1}{2}} \mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{\mu_2(\mathrm{d}\mathbf{x}^0)}{\sum_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2 \lambda_i}\right]^{\frac{1}{2}}, \qquad (90) \\
&= \left(\frac{1}{R_1}\sum_i \lambda_i\right)^{\frac{1}{2}} \mathbb{E}_{\mathrm{c},\theta^1}\left[\frac{1}{\sum_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2 \lambda_i}\right]^{\frac{1}{2}}, \\
&\geq \left(\frac{1}{R_1}\sum_i \lambda_i\right)^{\frac{1}{2}} \left(\mathbb{E}_{\mathrm{c},\theta^1}\left[\sum_i \left(\tilde{\boldsymbol{W}}^1_{\mathrm{c},i}\right)^2 \lambda_i\right]^{-1}\right)^{\frac{1}{2}} = 1. \qquad (91)
\end{aligned}$$

where Eq. (90) is 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) is 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

$$\begin{aligned}
\Theta^{l,h} &\equiv (\boldsymbol{\omega}^{1,1}, \boldsymbol{\beta}^{1,1}, \ldots, \boldsymbol{\omega}^{1,H}, \boldsymbol{\beta}^{1,H}, \ldots, \boldsymbol{\omega}^{l,1}, \boldsymbol{\beta}^{l,1}, \ldots, \boldsymbol{\omega}^{l,h}, \boldsymbol{\beta}^{l,h}), \qquad \theta^{l,h} \equiv \Theta^{l,h}|\Theta^{l,h-1}, \\
\Theta^{l} &\equiv (\boldsymbol{\omega}^{1,1}, \boldsymbol{\beta}^{1,1}, \ldots, \boldsymbol{\omega}^{1,H}, \boldsymbol{\beta}^{1,H}, \ldots, \boldsymbol{\omega}^{l,1}, \boldsymbol{\beta}^{l,1}, \ldots, \boldsymbol{\omega}^{l,H}, \boldsymbol{\beta}^{l,H}), \qquad \theta^{l} \equiv \Theta^{l}|\Theta^{l-1}.
\end{aligned}$$

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 ϕ are completely deterministic conditionally on Θ^{l-1} in the first layer $h = 1$ of each residual unit l. This occurs again for $h \geq 2$ since BN and ϕ are random conditionally on Θ^{l-1} but completely deterministic conditionally on $\Theta^{l,h-1}$. At even larger granularity, due to the aggregation $(\mathbf{y}^l, \mathrm{d}\mathbf{y}^l) = \sum_{k=0}^{l}(\mathbf{y}^{k,H}, \mathrm{d}\mathbf{y}^{k,H})$, 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 randomness θ^{l-k} of previous individual units.

Since the evolution of χ^l is mainly influenced by batch normalization and the nonlinearity ϕ, this shift can be thought as attributing the parameters and thus the stochasticity of layer h to layer $h-1$. A simple strategy to apply the results of Section 6 is thus to shift back to the post-activation perspective by considering the parameters $\theta^{l,h-1}$ and the evolution from $(\mathbf{x}^{l,h-1}, \mathrm{d}\mathbf{x}^{l,h-1})$ 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 Θ^{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, symmetric propagation with respect to Θ^{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 $\exp\left(\overline{m}_{\phi}[\chi^l]\right)$ 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

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

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

F.2. Lemma on dot-product

Lemma 14. *We have the following:*

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

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

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

Next we note that

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

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

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

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

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

where the dot product in Eq. (93) is computed in the orthogonal basis $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_N)$. Now computing the dot product of $\hat{\varphi}(\mathbf{y}^{l-1}, \boldsymbol{\alpha})$ and $\hat{\varphi}(\mathbf{y}^{l,H}, \boldsymbol{\alpha})$ in the orthogonal basis $(\boldsymbol{e}_1, \dots, \boldsymbol{e}_N)$:

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

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

$$\begin{aligned}
&\forall j \neq i:\ \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_i u_i\big]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_j u_j\big] \sim_{\theta^l} \mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[-y_i u_i\big]\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha}}\big[y_j u_j\big],\\
&\forall j \neq i:\ \alpha_{i,i}\alpha_{j,j} \sim_{\theta^l} (-\alpha_{i,i})\alpha_{j,j},\\
&\forall j \neq i:\ \mathbb{E}_{\theta^l}\big[\alpha_{i,i}\alpha_{j,j}\big] = 0.
\end{aligned}$$

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 θ^l does not depend on i. Denoting (β_j) such that $\forall i, j\colon \beta_j \equiv \mathbb{E}_{\theta^l}\big[\alpha_{i,j}^2\big]$, we get combined with Eq. (93):

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

Finally combining with Eq. (92):

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

where we used $\lambda_{\max} r_{\mathrm{eff}}(\mathbf{y}^{l-1}) = \sum_i \lambda_i = N\mu_2(\mathbf{y}^{l-1})$. The same analysis can be immediately transposed 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 as*

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

Then by spherical symmetry,

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

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

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

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*

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

Proof. First we introduce the additional constants $\gamma_{\min} \equiv (\delta_{\min})^{2H}$ and $\gamma_{\max} \equiv (\delta_{\max})^{2H}$, so 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:

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

Finally let us consider a random variable a depending on Θ^l with well-defined moments, and a constant b. 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 we had non negligible probability with respect to Θ^{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 $\mathbb{P}_{\Theta^l}[A^c] = \mathbb{E}_{\Theta^{l-1}}\mathbb{E}_{\Theta^l|\Theta^{l-1}}[\mathbf{1}_{A^c}] = \mathbb{E}_{\Theta^{l-1}}\mathbb{P}_{\Theta^l|\Theta^{l-1}}[A^c] \ll 1$. It follows that $\mathbb{P}_{\theta^l}[A^c] \ll 1$ with high probability with respect to Θ^{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

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

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

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

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

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

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

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

On the other hand, Corollary 15 implies that

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

Further using Chebyshev's inequality, we deduce that

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

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

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

Eq. (98) then holds for all l and we have $\big|Y_l\big| \ll \mu_2(\mathbf{y}^{l-1})$ 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

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

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

$$\tilde{T}_{l,l} = \frac{\mu_2(\mathbf{y}^0)}{\mu_2(\mathrm{d}\mathbf{y}^0)(\chi^{l-1})^2}\mu_2(\mathrm{d}\mathbf{y}^{l,H}) = \frac{\mu_2(\mathbf{y}^0)}{\mu_2(\mathrm{d}\mathbf{y}^0)(\chi^{l-1})^2}(\chi^{l-1})^2\prod\nolimits_h(\delta\chi^{l,h})^2\mu_2(\mathbf{y}^{l,H})\frac{\mu_2(\mathrm{d}\mathbf{y}^0)}{\mu_2(\mathbf{y}^0)},$$
$$\gamma_{\min}\mu_{2,\min} \lesssim \tilde{T}_{l,l} \lesssim \gamma_{\max}\mu_{2,\max}. \tag{100}$$

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

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

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

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

□

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

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

We can further explicitate the bounds:

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

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

$$\sum_{k=1}^{l}\log\left(1 + \frac{\eta_{\min}}{k+1}\right)$$

$$\geq \eta_{\min}\log(l+2+\eta_{\min}) + (l+2)\log\left(1+\frac{\eta_{\min}}{l+2}\right) - (2+\eta_{\min})\log(2+\eta_{\min}) + 2\log 2,$$
$$\geq \eta_{\min}\log(l+2+\eta_{\min}) - (2+\eta_{\min})\log(2+\eta_{\min}) + 2\log 2.$$

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

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

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

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

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

F.4. Theorem 4 holds for any choice of ϕ with and without batch normalization, as long as the existence of $\mu_{2,\min}$, $\mu_{2,\max}$, $\delta_{\min}$, $\delta_{\max}$ is ensured

The proof of Lemma 14 neither requires batch normalization, nor does it require any assumption on ϕ. 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

$$\mathbb{E}_{\theta^l}\left[\mathbb{E}_{\mathbf{y},\boldsymbol{\alpha},\mathrm{c}}\Big[\varphi(\mathbf{y}^{l-1},\boldsymbol{\alpha})_{\mathrm{c}}\varphi(\mathbf{y}^{l,H},\boldsymbol{\alpha})_{\mathrm{c}}\Big]^2\right] \leq \frac{1}{N^2}\lambda_{\max}\mathbb{E}_{\theta^l}\left[\nu_2(\mathbf{y}^{l,H})\right] \tag{102}$$
$$\leq \frac{1}{N}\nu_2(\mathbf{y}^{l-1})\mathbb{E}_{\theta^l}\left[\nu_2(\mathbf{y}^{l,H})\right]. \tag{103}$$

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

It is now interesting to determine when $\mu_{2,\min}$, $\mu_{2,\max}$, $\delta_{\min}$, $\delta_{\max}$ 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 $\mu_{2,\min}$, $\mu_{2,\max}$ is ensured.

Now let us see whether δ_{min}, δ_{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 expectation over Θ^l equivalent to averaging over channels. Using Lemma 14 as well as Eq. (103), 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 ϕ 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, meaning that the existence of δ_{max} is *not* ensured. Now if we replace η_{min}, η_{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 deduce that $\chi^l = \prod_k \delta\chi^k = \exp\left(\sum_k \log \delta\chi^l\right)$ scales as $\exp\left(\frac{\mathcal{A}}{2}\sqrt{l}\right)$. Combined with $\mu_2(\mathbf{y}^{l-1}) \propto l$ and the definition of χ^l, we deduce that $\mu_2(\mathrm{d}\mathbf{y}^{l-1})$ scales as $\exp\left(\mathcal{A}\sqrt{l}\right)$, which is exactly the scaling found in Yang & Schoenholz (2017) for the corresponding quantity.

In summary, the growth of χ^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 ϕ step and a single convolution step, the existence of $\mu_{2,\text{min}}$, $\mu_{2,\text{max}}$ is ensured.

Now let us see whether δ_{min}, δ_{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 ϕ. With the assumption of well-conditioned noise, we thus deduce that $\exp(\overline{m}_{BN}[\chi^l])$ 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: $N \to \infty$, where $\mathbf{z}^l$ 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 χ^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, we deduce that moments $\nu_2(\mathbf{y}^{l,h})$, $\mu_2(\mathrm{d}\mathbf{y}^{l,h})$ are roughly stable and that increments $\delta\chi^{l,h}$ are limited: $\delta\chi^{l,h} \simeq 1$ 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 that $\mu_2(\mathbf{y}^{l,H}) \simeq \mu_2(\mathbf{y}^{l,0}) = \mu_2(\mathbf{y}^{l-1})$. 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. (103), 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 $\mu_{2,\max}$ 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 $\delta_{\min}$, $\delta_{\max}$ 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, it follows that $\delta\chi^{l,h}$ is bounded from above and below.

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

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

As in Section F.4.2, the existence of $\mu_{2,\min}$, $\mu_{2,\max}$ 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 ϕ step and a single convolution step.

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

Theorem 3 further directly 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, it follows that $\delta\chi^{l,h}$ is bounded from above and below.

In summary, Theorem 4 applies and χ^l has power-law growth.