Characterizing Well-Behaved vs. Pathological Deep Neural Networks

Antoine Labatie [1]

Abstract

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

1. Introduction

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

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

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

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

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

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

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

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

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

2. Propagation

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

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

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

with $*$ the convolution and $\boldsymbol{\beta}^l \in \mathbb{R}^{n\times\cdots\times n\times N_l}$ the tensor with repeated version of $\mathbf{b}^l$ at each spatial position. From now on, we refer to the propagated tensor $\mathbf{x}^l$ as the *signal*.

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

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

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

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

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

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

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

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

3. Data Randomness

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

3.1. Characterizing Data Distributions

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

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

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

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

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

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

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

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

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

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

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

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

with $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}$ the covariance matrix and $||\cdot||$ the spectral norm. If we further denote (λ_i) the eigenvalues of $\boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\boldsymbol{\alpha}}[\varphi(\mathbf{v}^l,\boldsymbol{\alpha})]$, then $r_{\mathrm{eff}}(\mathbf{v}^l) = \sum_i \lambda_i / \max_i \lambda_i \geq 1$. Intuitively, $r_{\mathrm{eff}}(\mathbf{v}^l)$ 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, i.e. 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 $\chi^l > 1$ are noise amplifiers, while neural networks with $\chi^l < 1$ are noise reducers.

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

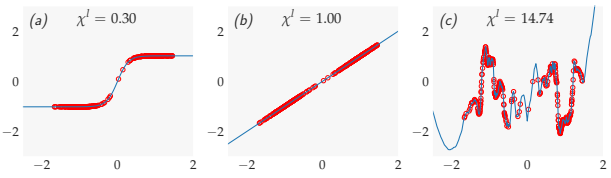

Figure 1: *Illustration of* χ^l in the fully-connected case with one-dimensional input and output, $N_0 = 1$, $N_l = 1$. We show the full input-output mapping Φ_l *(blue curves)* and randomly sampled input-output data points $\big(\mathbf{x}^0, \Phi_l(\mathbf{x}^0)\big)$ *(red circles)* for three different neural networks sharing the same input signal $\mathbf{x}^0$ and the same variance in their output signal $\mu_2(\Phi_l(\mathbf{x}^0))$. *(a)* Since input data points $\mathbf{x}^0$ appear in flat regions of 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 outlined, χ^l measures the sensitivity to signal perturbation, which is known for being connected to generalization (Rifai et al., 2011; Arpit et al., 2017; Sokolic et al., 2017; Arora et al., 2018; Morcos et al., 2018; Novak et al., 2018; Philipp & Carbonell, 2018). A tightly connected notion is the sensitivity to weight perturbation, also known for being connected to generalization (Hochreiter & Schmidhuber, 1997; Langford & Caruana, 2002; Keskar et al., 2017; Chaudhari et al., 2017; Smith & Le, 2018; Dziugaite & Roy, 2017; Neyshabur et al., 2017; 2018; Li et al., 2018). The connection is seen by noting the equivalence between a noise $\mathrm{d}\boldsymbol{\omega}^l$ on the weights and a noise $\mathrm{d}\mathbf{y}^l = \mathrm{d}\boldsymbol{\omega}^l * \mathbf{x}^{l-1}$ and $\mathrm{d}\mathbf{x}^l = \phi'(\mathbf{y}^l) \odot \mathrm{d}\mathbf{y}^l$ on the signal in Eq. (1) and Eq. (2).

3.2. Characterizing Pathologies

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

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

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

The pathology $\mu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^l) \to 0$ implies $\mu_2(\tilde{\mathbf{x}}^l) \to 0$, meaning that $\varphi(\tilde{\mathbf{x}}^l, \boldsymbol{\alpha})$ becomes point-like concentrated at the point $\tilde{\boldsymbol{\nu}}^l$ of unit L^2 norm: $||\tilde{\boldsymbol{\nu}}^l||_2 = 1$ [C.4]. In the limit of strict point-like concentration, the upper neural network from layer l to L is limited to random guessing since it "sees" all inputs the same and cannot distinguish between them.

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

– *Exploding sensitivity*: $\chi^l \geq \exp(\gamma l) \xrightarrow{l\to\infty} \infty$ for some $\gamma > 0$. Given the noise factor equivalence of Eq. (4), the pathology $\chi^l \to \infty$ implies $\text{SNR}^l \to 0$, meaning that the clean signal $\mathbf{x}^l$ becomes drowned in the noise $\mathrm{d}\mathbf{x}^l$. In the limit of strictly zero signal-to-noise ratio, the upper neural network from layer l to L is limited to random guessing since it only "sees" noise.

4. Model Parameters Randomness

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

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

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

– In the case of zero-dimensional signal, the upper neural network from layer l to L must adjust its bias parameters very precisely in order to center the signal and distinguish between different inputs. This case – further associated with vanishing gradients for bounded 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 $\phi = \text{ReLU}$ 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).

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 the model parameters as $\Theta^l \equiv (\boldsymbol{\omega}^1, \boldsymbol{\beta}^1, \ldots, \boldsymbol{\omega}^l, \boldsymbol{\beta}^l)$ and use θ^l as shorthand for $\Theta^l | \Theta^{l-1}$. We further denote the geometric increments of $\nu_2(\mathbf{x}^l)$ as $\delta\nu_2(\mathbf{x}^l) \equiv \nu_2(\mathbf{x}^l)/\nu_2(\mathbf{x}^{l-1})$.

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

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

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

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

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

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

Discussion. We directly note that: (i) $\overline{m}[\nu_2(\mathbf{x}^k)]$ and $\underline{m}[\nu_2(\mathbf{x}^k)]$ are random variables which depend on 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 $\nu_2(\mathbf{x}^k) = \frac{1}{N_k}\sum_c \nu_{2,c}(\mathbf{x}^k)$, 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 the *layer-wise* character of this approximation, whose validity only requires $N_l \gg 1$, independently of the depth l. This contrasts with the *aggregated* character (up to layer l) of the mean field approximation of $\mathbf{y}^l$ as a Gaussian process, whose validity requires not only $N_l \gg 1$ but also – as we will see – that the depth l remains sufficiently small with respect to N_l.

5. Vanilla Nets

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

Theorem 1 (moments of vanilla nets)**.** [D.3] *There exist small constants* $1 \gg m_{\min}, m_{\max}, v_{\min}, v_{\max} > 0$*, 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 probabilities of the complementary events A_l^c, $A_l'^c$ decay exponentially in the width N_l [D.4].

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

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

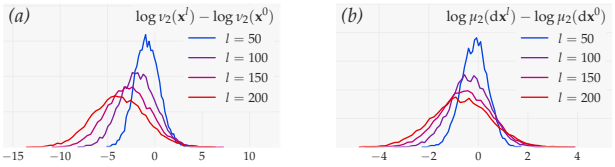

Figure 2: *Slowly diffusing moments of vanilla nets* with $L = 200$ layers of width $N_l = 128$. *(a)* Distribution of $\log \nu_2(\mathbf{x}^l) - \log \nu_2(\mathbf{x}^0)$ for $l = 50, 100, 150, 200$. *(b)* Same for $\log \mu_2(\mathrm{d}\mathbf{x}^l) - \log \mu_2(\mathrm{d}\mathbf{x}^0)$.

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

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

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

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

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

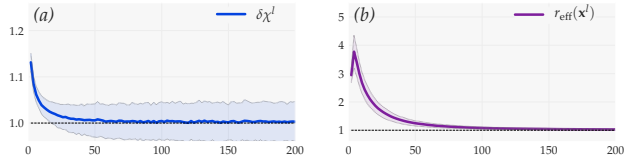

Figure 3: *Pathology of one-dimensional signal for vanilla nets* with $L = 200$ layers of width $N_l = 512$. *(a)* $\delta\chi^l$ such that $\delta\chi^l \simeq \exp\left(\overline{m}[\chi^l]\right) \to 1$. *(b)* $r_{\text{eff}}(\mathbf{x}^l)$ indicates one-dimensional signal pathology: $r_{\text{eff}}(\mathbf{x}^l)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

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 is such that $\exp(\overline{m}_{\text{BN}}[\chi^l]) \simeq \delta_{\text{BN}}\chi^l$, with $\delta_{\text{BN}}\chi^l$ defined as the increment of 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 expectation 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 the magnitude of $r_{\text{eff}}(\mathbf{x}^{l-1})$, $r_{\text{eff}}(\mathrm{d}\mathbf{x}^{l-1})$. If we assume that $\mathrm{d}\mathbf{x}^{l-1}$ is well-conditioned, then $\mu_{2,\text{c}}(\mathrm{d}\mathbf{y}^l)/\mu_2(\mathrm{d}\mathbf{x}^{l-1})$ can be treated as a constant and by convexity of the function $x \mapsto 1/x$:

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

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 wherein the noise has perfect conditioning, resulting in $\exp(\overline{m}_{\text{BN}}[\chi^1]) \geq 1$ [E.2].

Effect of the Nonlinearity. The nonlinearity term is such that $\exp(\overline{m}_\phi[\chi^l]) \simeq \delta_\phi\chi^l$, with $\delta_\phi\chi^l$ defined as the increment of 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: (i) $\mathrm{d}\mathbf{x}^l$ remains well-conditioned, while $\mathbf{x}^l$ becomes ill-conditioned; (ii) $r_{\text{eff}}(\mathbf{x}^l)$ and $\delta_{\text{BN}}\chi^l$ are inversely correlated.

Interestingly, $\delta_\phi\chi^l$ becomes subdominant with respect to $\delta_{\text{BN}}\chi^l$ at large depth. This stems from the fact that $\mathbf{z}^l$ becomes fat-tailed distributed with respect to $\mathbf{x}$, $\boldsymbol{\alpha}$, with large $\mu_4(\mathbf{z}^l)$ and small $\nu_1(|\mathbf{z}^l|)$. Combined with $\nu_1(\mathbf{z}^{l,+}) \leq \nu_1(|\mathbf{z}^l|)$ and $\nu_1(\mathbf{z}^{l,-}) \leq \nu_1(|\mathbf{z}^l|)$, this explains the decay of $|\exp(\overline{m}_\phi[\chi^l]) - 1|$ and thus of $|\delta_\phi\chi^l - 1|$.

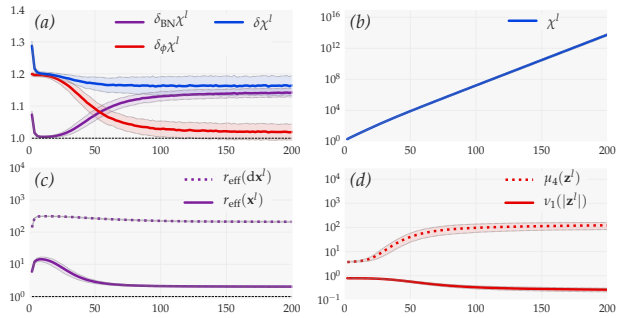

Figure 4: *Pathology of exploding sensitivity for batch-normalized feedforward nets* with $L = 200$ layers of width $N_l = 512$. *(a)* Geometric increments $\delta\chi^l$ decomposed as the product of $\delta_{\text{BN}}\chi^l$ defined as the increment from $(\mathbf{x}^{l-1}, \mathrm{d}\mathbf{x}^{l-1})$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

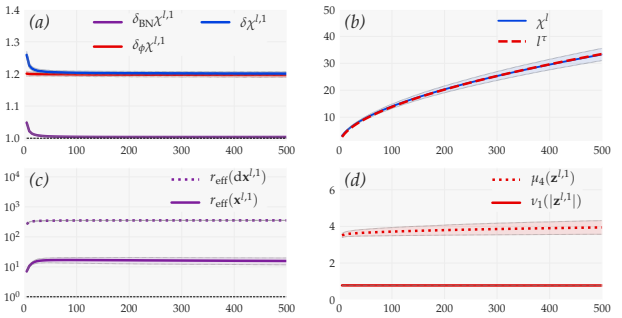

Figure 5: *Well-behaved evolution of batch-normalized resnets* with $L = 500$ residual units comprised of $H = 2$ layers of width $N = 512$. *(a)* Geometric feedforward increments $\delta\chi^{l,1}$ decomposed as the product of $\delta_{\text{BN}}\chi^{l,1}$ defined as the increment from $(\mathbf{y}^{l,0},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

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

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

Theorem 4 (normalized sensitivity increments of batch-normalized resnets)**.** [F.3] *Suppose that we can bound signal variances:* $\mu_{2,\min} \lesssim \mu_2(\mathbf{y}^{l,H}) \lesssim \mu_{2,\max}$ *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: $\mu_{2,\mathrm{c}}(\mathbf{z}^{l,H}) = 1$ [F.4].

Now let us get a better grasp of Theorem 4. We see in Eq. (14) that the evolution remains exponential inside residual units since $\eta_{\min}$, $\eta_{\max}$ have an exponential dependence in H. However, it is slowed down by the factor $1/(l+1)$ between successive residual units. This stems from the dilution (Philipp et al., 2018) of the residual path $(\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H})$ into the skip-connection path $(\mathbf{y}^{l-1}, \mathrm{d}\mathbf{y}^{l-1})$ with ratio of signal variances: $\mu_2(\mathbf{y}^{l,H})/\big(\mu_2(\mathbf{y}^{l,H}) + \mu_2(\mathbf{y}^{l-1})\big)$ decaying as $1/(l+1)$. If we remove the dilution effect by multiplying the residual branch by 0 (i.e. replacing the scaling in $1/(l+1)$ by

a scaling in 1) and if we set $\mu_{2,\text{min}} = \mu_{2,\text{max}}$, then Eq. (14) recovers the feedforward evolution $(\delta_{\text{min}})^H \lesssim \delta\chi^l \lesssim (\delta_{\text{max}})^H$. The dilution is clearly visible in Eq. (13). Namely, each residual unit adds a term $(\mathbf{y}^{l,H}, \mathrm{d}\mathbf{y}^{l,H})$ of increased $\chi^{l,H}$ but its relative contribution to the aggregation gets smaller and smaller with l, so that the growth of 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) as

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

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

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

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

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

– In the context of vanilla nets, with e.g. an input $\varphi(\mathbf{x}, \boldsymbol{\alpha})$ constant with respect to $\boldsymbol{\alpha}$ and reduced to a single point of $\mathbb{R}^{N_0}$ such that $\varphi(\mathbf{x}^l, \boldsymbol{\alpha})$ remains a single point of $\mathbb{R}^{N_l}$. Given the evolution of Fig. 2, the L^2 norm $||\varphi(\mathbf{x}^l, \boldsymbol{\alpha})||_2^2 = N_l \nu_2(\mathbf{x}^l)$ becomes fat-tailed distributed as $l \to \infty$. For given $\mathbf{x}, \boldsymbol{\alpha}, \mathrm{c}$, this means that $\mathbf{x}^l_{\boldsymbol{\alpha},\mathrm{c}}$ and thus $\mathbf{y}^l_{\boldsymbol{\alpha},\mathrm{c}}$ become fat-tailed distributed as $l \to \infty$.

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

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

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

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

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

– 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.

Acknowledgements

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

References

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

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

Balduzzi, D., Frean, M., Leary, L., Lewis, J. P., Ma, K. W., and McWilliams, B. The shattered gradients problem: If resnets are the answer, then what is the question? In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 342–350, 2017. URL http://proceedings.mlr.press/v70/balduzzi17b.html.

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

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

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

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

Duvenaud, D., Rippel, O., Adams, R., and Ghahramani, Z. Avoiding pathologies in very deep networks. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 202–210, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL http://proceedings.mlr.press/v33/duvenaud14.html.

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

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

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

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

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pp. 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.123. URL http://dx.doi.org/10.1109/ICCV.2015.123.

He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 630–645, 2016. doi: 10.1007/

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

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

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

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

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

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

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

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

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

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

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

Neyshabur, B., Bhojanapalli, S., Mcallester, D., and Srebro, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems 30*, pp. 5947–5956. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7176-exploring-generalization-in-deep-learning.pdf.

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

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

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

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

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

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

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S., and Sohl-Dickstein, J. On the expressive power of deep neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2847–2854, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/raghu17a.html.

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

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

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

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

Sokolic, J., Giryes, R., Sapiro, G., and Rodrigues, M. R. D. Robust large margin deep neural networks. *IEEE Trans. Signal Processing*, 65(16):4265–4280, 2017. doi: 10.1109/TSP.2017.2708039. URL https://doi.org/10.1109/TSP.2017.2708039.

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

Williams, C. K. I. Computing with infinite networks. In *Advances in Neural Information Processing Systems 9*, pp. 295–301. MIT Press, 1997. URL http://papers.nips.cc/paper/1197-computing-with-infinite-networks.pdf.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

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

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

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