Characterizing Well-Behaved vs. Pathological Deep Neural Networks

Antoine Labatie (ICML 2019)

Context

There is still no **mature theory** able to validate the **full choice of hyperparameters** leading to state-of-the-art performance in deep neural networks.

A large branch of research aimed at building this theory has focused on networks at the time of **random initialization**. The justification is twofold:

1. Due to the randomness of model parameters at initialization, networks at that time may serve as a proxy for the full hypothesis space
2. The initialization has an importance in itself as the starting point of the optimization

Our contributions:

1. We introduce a **unifying methodology** to characterize neural networks at initialization
2. We apply this methodology to characterize neural networks with the most commonly used hyperparameters

Methodology — Propagation

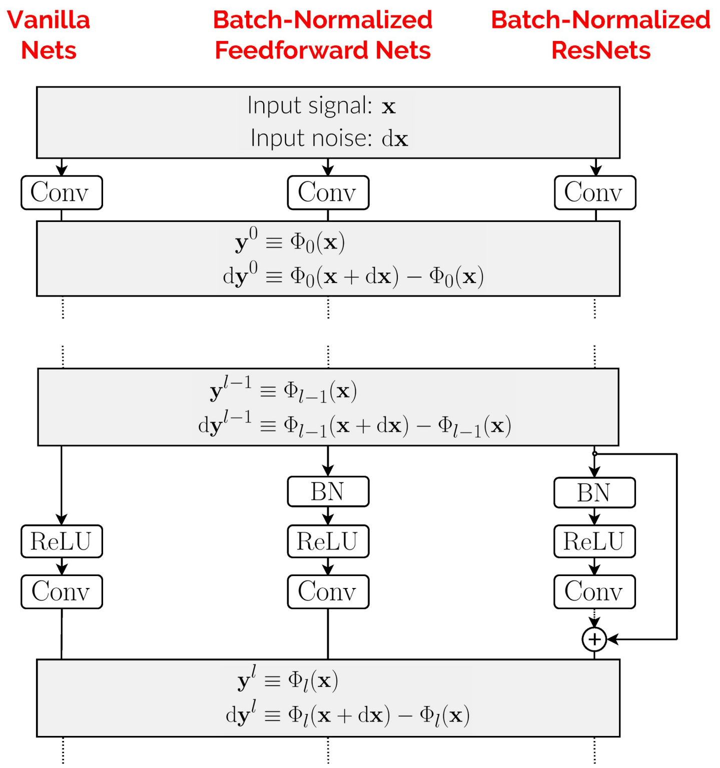
Simultaneous propagation of:

- The **signal**
- An **additive noise** corrupting the signal

Vanilla Nets:
convolution + ReLU

Batch-Normalized Feedforward Nets:
convolution + batch norm + ReLU

Batch-Normalized ResNets:
convolution + batch norm + ReLU + skip connection



Methodology — Data Randomness

Effective Rank:

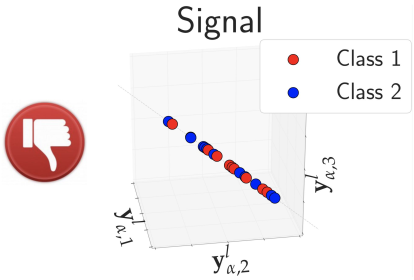
$$r_{\text{eff}}(\mathbf{y}^l) \equiv \frac{\operatorname{Tr} \boldsymbol{C}_{\mathbf{x},\alpha}\left[\mathbf{y}^l_{\alpha,:}\right]}{\left|\left|\boldsymbol{C}_{\mathbf{x},\alpha}\left[\mathbf{y}^l_{\alpha,:}\right]\right|\right|} = \frac{\sum_i \lambda_i}{\max_i \lambda_i} \geq 1.$$

Normalized Sensitivity:

$$\chi^l \equiv \left(\frac{\mathrm{SNR}^0}{\mathrm{SNR}^l}\right)^{\frac{1}{2}}, \quad \text{with } \mathrm{SNR}^l \equiv \frac{\operatorname{Tr} \boldsymbol{C}_{\mathbf{x},\alpha}\left[\mathbf{y}^l_{\alpha,:}\right]}{\operatorname{Tr} \boldsymbol{C}_{\mathbf{x},\mathrm{d}\mathbf{x},\alpha}\left[\mathrm{d}\mathbf{y}^l_{\alpha,:}\right]}.$$

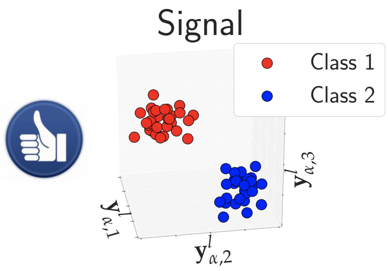
Pathology of One-Dimensional Signal:

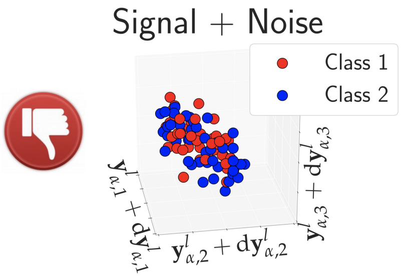
$$r_{\text{eff}}(\mathbf{y}^l) \xrightarrow{l \to \infty} 1.$$



Pathology of Exploding Sensitivity:

$$\chi^l \geq \exp(\gamma l) \xrightarrow{l \to \infty} \infty, \quad \text{for some } \gamma > 0.$$





Methodology — Model Parameters Randomness

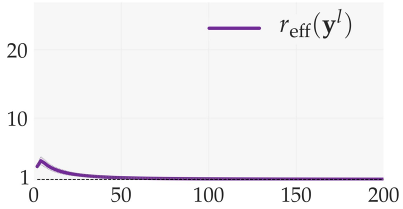
Finally, we introduce the randomness from **model parameters** at **initialization**.

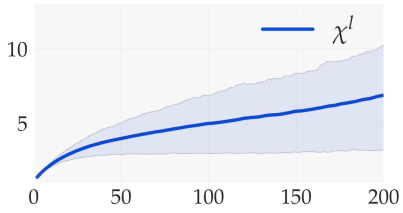
The key of our methodology consists in treating the effective rank and the normalized sensitivity as **random variables** which depend on these **model parameters.**

Applying the Methodology

Vanilla Nets

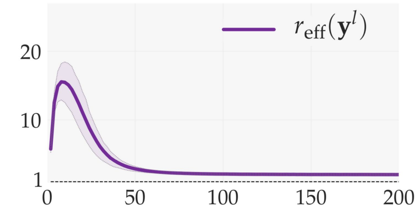
- Pathology of one-dimensional signal
- Limited growth of the normalized sensitivity

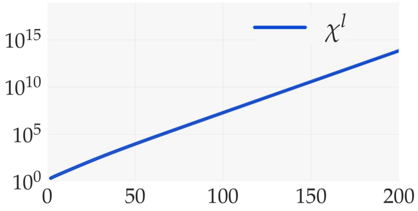




Batch-Normalized Feedforward Nets

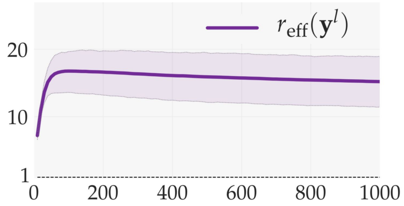
- Few directions of signal variance preserved
- Pathology of exploding sensitivity

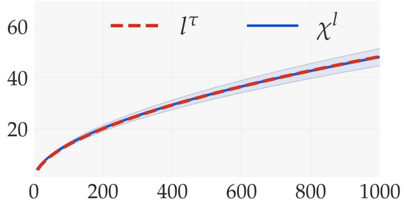




Batch-Normalized ResNets

- Many directions of signal variance preserved
- Power-law growth of the normalized sensitivity





Takeaway

There are two opposing forces at work:

1. The **additivity** of convolutions (i.e. affine transforms) with respect to width, which repels from pathologies
2. The **multiplicativity** of layer composition with respect to depth, which attracts to pathologies

Feedforward nets are pathological at high depth since they are subject both to additivity and multiplicativity.

Batch-normalized resnets are well-behaved at all depths since they are subject to additivity but relieved from multiplicativity.

Characterizing Well-Behaved vs. Pathological Deep Neural Networks

Pacific Ballroom #98

Code: https://github.com/alabatie/moments-dnns