Characterizing Well-Behaved vs. Pathological Deep Neural Networks

Antoine Labatie (ICML 2019)

Context

Deep neural networks have been tremendously successful in many applications. Yet, there is still a lack of a *mature theory* able to validate the *full choice of hyperparameters* associated with state-of-the-art performance.

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

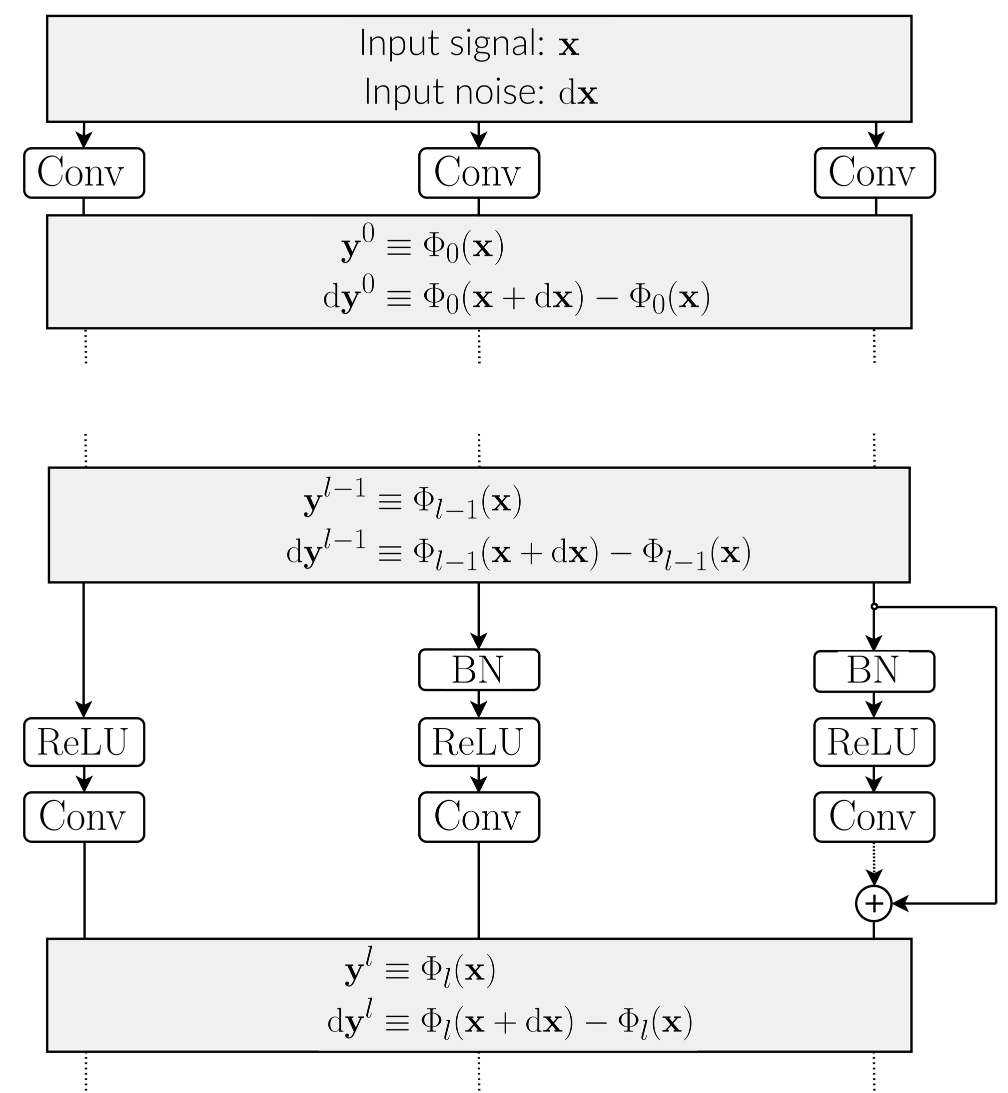
1. Due to the randomness of model parameters at initialization, networks at that time may serve as a proxy for the full hypothesis space
2. Pathologies in neural networks at initialization are likely – in any case – to penalize training by hindering optimization

Contributions

1. We introduce a novel approach to characterize deep neural networks at initialization:
   - the treatment of the broad spectrum of pathologies is unified
   - only mild assumptions are required
   - convolutional layers, batch normalization, skip connections are easily incorporated
2. Using this approach, we characterize deep neural networks with the most common choices of hyperparameters

Propagation

Vanilla Nets	Batch-Normalized Feedforward Nets	Batch-Normalized ResNets



Methodology — Data Randomness

First, we introduce the *data randomness* coming from the input signal $\mathbf{x}$, the input noise $\mathrm{d}\mathbf{x}$, and – in the convolutional case – the spatial position α. At this point, model parameters are *fixed*.

Effective Rank — Pathology of One-Dimensional Signal

The **effective rank** is defined as:

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

with $\boldsymbol{C}_{\mathbf{x},\alpha}\big[\mathbf{y}^l_{\alpha,:}\big]$ the covariance matrix of $\mathbf{y}^l_{\alpha,:}$ and (λ_i) its eigenvalues.

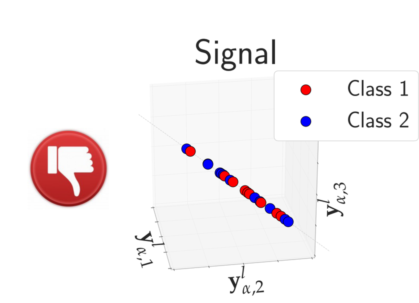
$r_{\mathrm{eff}}(\mathbf{y}^l)$ measures the number of *effective directions* which concentrate the variance of $\mathbf{y}^l_{\alpha,:}$.

The **pathology of one-dimensional signal** is characterized by:

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

This pathology implies that $\mathbf{y}^l_{\alpha,:}$ becomes *line-like* concentrated.

The consequence is that layers $l' > l$ only "see" a single feature from the signal.



Normalized Sensitivity — Pathology of Exploding Sensitivity

The **normalized sensitivity** is defined as:

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

Neural networks with $\chi^l > 1$ *degrade the signal-to-noise ratio*, i.e. are *noise amplifiers*.

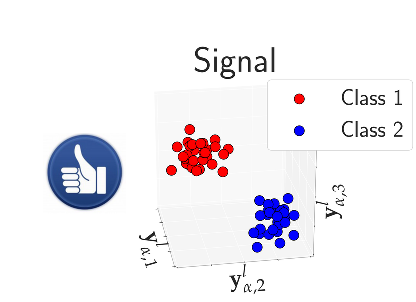
Neural networks with $\chi^l < 1$ *enhance the signal-to-noise ratio*, i.e. are *noise reducers*.

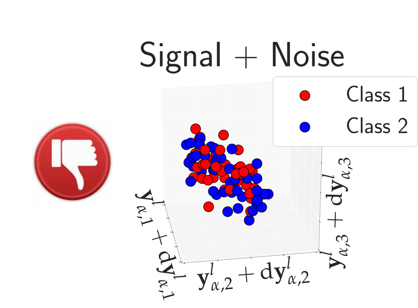
The **pathology of exploding sensitivity** is characterized by:

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

This pathology implies that $\mathbf{y}^l_{\alpha,:}$ becomes *drowned in the noise* $\mathrm{d}\mathbf{y}^l_{\alpha,:}$.

The consequence is that layers $l' > l$ only "see" noise.





Methodology — Model Parameters Randomness

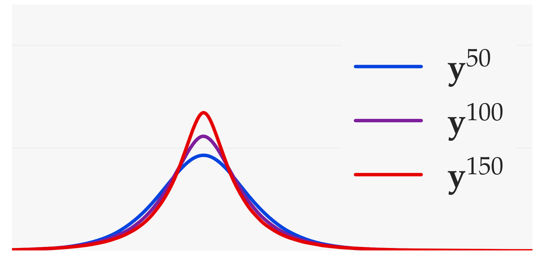
Second, we introduce the *model parameters randomness* at the time of random initialization. We suppose that the initialization is standard [1, 2].

The key of our methodology consists in treating $r_{\mathrm{eff}}(\mathbf{y}^l)$, χ^l as random variables which depend on model parameters.

Results — Gaussianity

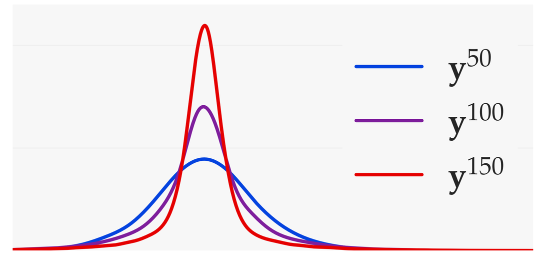
Vanilla Nets

For fixed data and random model parameters, the distribution of $\mathbf{y}^l$ *may become non-Gaussian* at high depth.



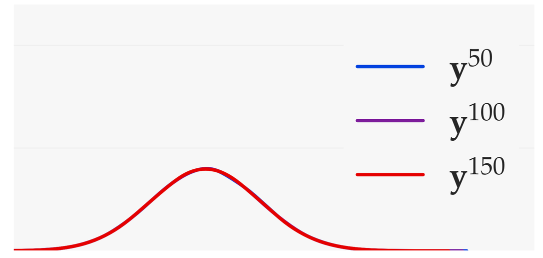
Batch-Normalized Feedforward Nets

For fixed data and random model parameters, the distribution of $\mathbf{y}^l$ *may become non-Gaussian* at high depth.



Batch-Normalized ResNets

For fixed data and random model parameters, the distribution of $\mathbf{y}^l$ *remains Gaussian* at all depths.



Results — Pathologies

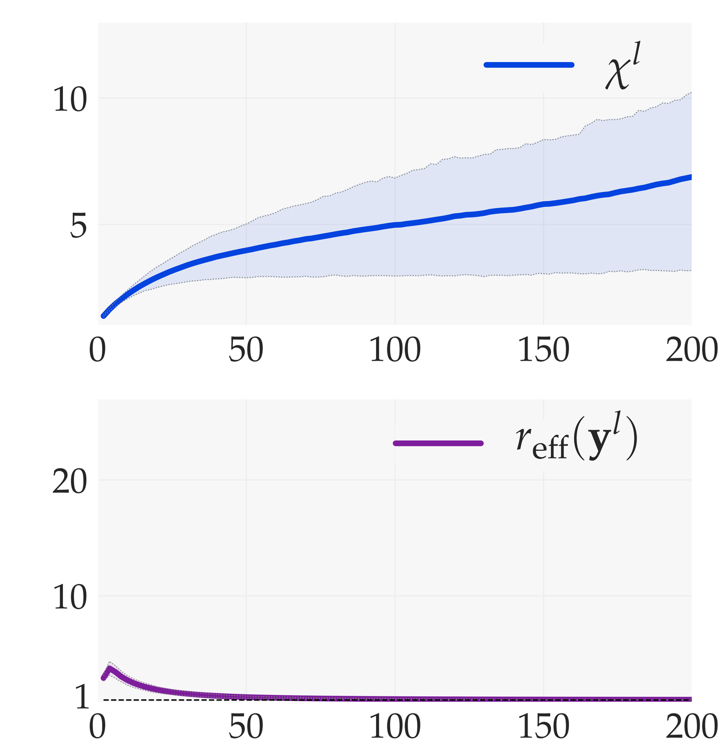
Vanilla Nets

Theory:

- $1 \lesssim \delta\chi^l \equiv \frac{\chi^l}{\chi^{l-1}} \lesssim \sqrt{2}$
- If $\delta\chi^l \xrightarrow{l\to\infty} 1$, then $r_{\mathrm{eff}}(\mathbf{y}^l) \xrightarrow{l\to\infty} 1$

Experiments:

- Limited growth of the normalized sensitivity: $\delta\chi^l \xrightarrow{l\to\infty} 1$
- *Pathology of one-dimensional signal*: $r_{\mathrm{eff}}(\mathbf{y}^l) \xrightarrow{l\to\infty} 1$



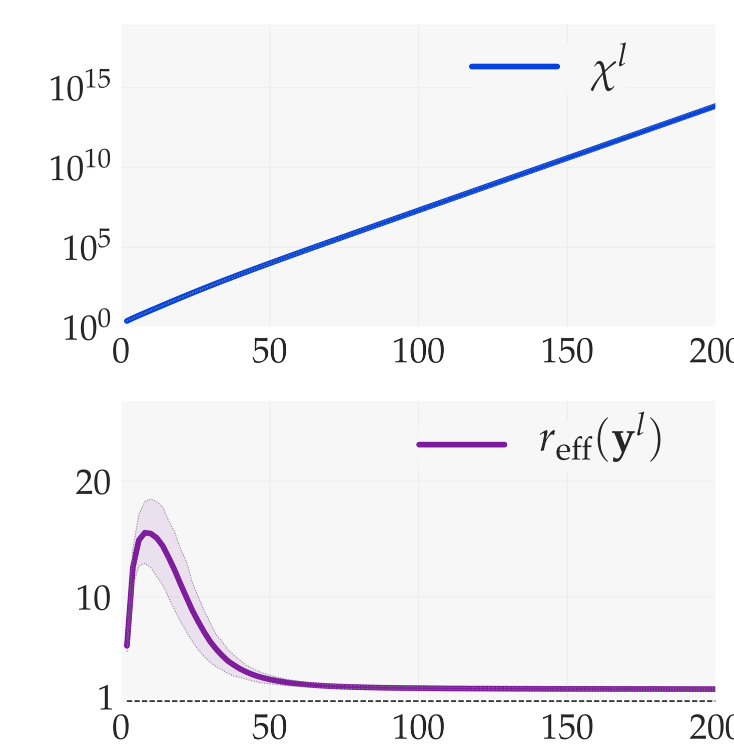
Batch-Normalized Feedforward Nets

Theory:

- $\delta\chi^l \equiv \frac{\chi^l}{\chi^{l-1}} \equiv \delta\chi^l_{\mathrm{BN}} \cdot \delta\chi^l_{\phi}$
- $1 \lesssim \delta\chi^l_{\mathrm{BN}}$ and $1 \lesssim \delta\chi^l_{\phi} \lesssim \sqrt{2}$

Experiments:

- *Pathology of exploding sensitivity*: $\chi^l \geq \exp(\gamma l) \xrightarrow{l\to\infty} \infty$, for some $\gamma > 0$
- Few directions of signal variance preserved in $r_{\mathrm{eff}}(\mathbf{y}^l)$



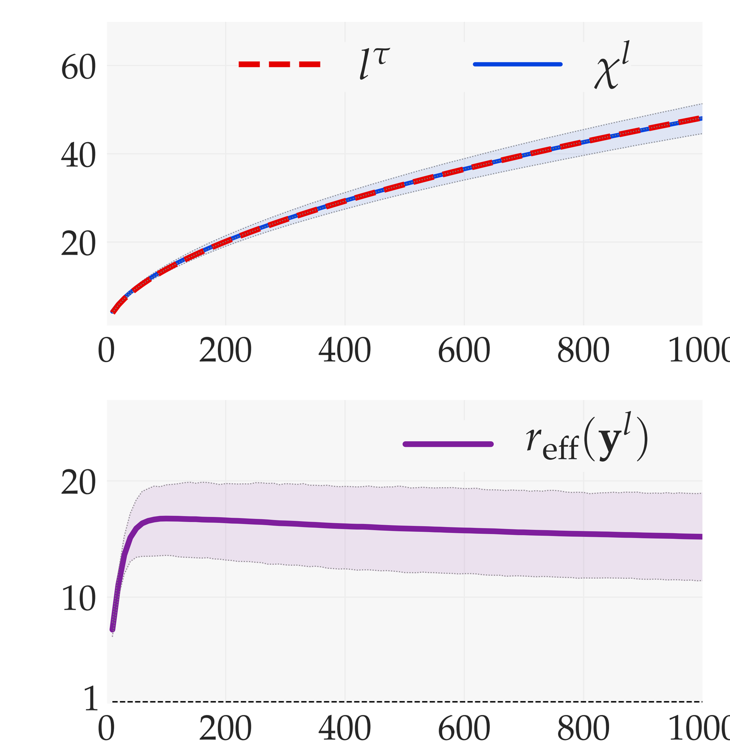
Batch-Normalized ResNets

Theory:

- $\left(1+\frac{\eta_{\min}}{l+1}\right)^{\frac{1}{2}} \lesssim \delta\chi^l \lesssim \left(1+\frac{\eta_{\max}}{l+1}\right)^{\frac{1}{2}}$
- $C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}$

Experiments:

- Power-law growth of the normalized sensitivity χ^l
- Many directions of signal variance preserved in $r_{\mathrm{eff}}(\mathbf{y}^l)$



Feedforward Nets are Pathological — Batch-Normalized ResNets are Well-Behaved

There are two opposing forces at work:

- The additivity with respect to width of affine transforms, which *repels from pathologies* and *attracts to Gaussianity*
- The multiplicativity with respect to depth of layer composition, which *attracts to pathologies* and *repels from Gaussianity*

Because they are *subject both to additivity and multiplicativity*, *feedforward nets are pathological* at high depth.
Because they are *subject to additivity but relieved from multiplicativity*, *batch-normalized resnets are well-behaved* at all depths.

Details of the Experiments

All our experiments were made with convolutional networks of width 512 on CIFAR-10.

alabatie/moments-dnns

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *ICCV 2015*.

[2] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ICML 2015*.