On the Inductive Bias of Various Architectures of Deep ReLU Nets (from ICML 2019)

Antoine Labatie

Context (I)

The recent years have seen huge advances in deep neural network (DNN) architectures, initialization, regularization.

In contrast, there is still no **mature theory** able to validate the **full choice of hyperparameters** leading to state-of-the-art performance.

This is unfortunate since such theory would provide a principled way of guiding:

1. A good choice of hyperparameters for new problems / setups
2. Further refinement in the choice of hyperparameters for known problems / setups

Context (II)

A large branch of research aimed at building this theory has focused on networks **at initialization**.

The justification is twofold:

1. Characterizing networks with randomly sampled model parameters amounts to characterizing the inductive bias of the architecture
2. The initialization has an importance in itself as the starting point of the optimization

In practice, “good properties” at initialization were found to be highly 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).

Related work (I)

This case of DNNs at initialization still involves difficulties

- Difficulty to deal with the complex interplay of the 2 sources of randomness:
  - The input data
  - The model parameters
- Difficulty to deal with the whole spectrum of potential “bad properties”
- Difficulty — paradoxically — to deal with the finite number of units in each layer
- Difficulty to incorporate hard-to-model components of state-of-the-art DNNs:
  - Convolutional layers
  - Batch normalization
  - Skip-connections
  - etc

Related work (II)

The most commonly used methodology is known as **“mean field”**:

- It assumes that DNNs are well approximated by Gaussian processes (limit of infinite width for any given depth)
  - The input data corresponds to the index of the Gaussian process
  - The model parameters correspond to the stochasticity of the Gaussian process
  - The Gaussian Process is fully described by its kernel
  - This approximation **breaks down** for feedforward nets at high depth
- It does not lead to straightforward definitions of “bad properties” of DNNs, so only specific cases of “bad properties” can be studied:
  - Exponential correlation / decorrelation of data points (Schoenholz et al., 2017; Xiao et al., 2018)
  - Exploding / vanishing gradients (Yang & Schoenholz, 2017; Yang et al., 2019)
- It does not allow an easy incorporation of conv layers, batch norm and skip-connections

Related work (III)

Other studies had similar limitations:

- Only specific cases of input data and pathologies were considered:
  - Exploding complexity of one-dimensional data manifolds (Raghu et al., 2017)
  - Exponential correlation / decorrelation of two data points (Balduzzi et al., 2017)
  - Exploding / vanishing gradients (Philipp et al., 2018)
- The hard-to-model components were either excluded or incorporated one at a time

In conclusion, all studies have been limited either in their **scope** or their **simplifying assumptions**.

Contributions

Two main contributions.

1. Introduction of a **novel methodology** to characterize **convolutional** and **fully-connected** DNNs at initialization:
    a. Offers a unifying treatment of the whole spectrum of “bad properties” without any restriction on the input data
    b. Requires only mild assumptions
    c. Easily incorporates conv layers, batch norm, skip-connections

1. Using this methodology, characterization of the inductive bias of various architectures of DNNs:
    a. Insights on the most suitable choices of architectures
    b. Insights on the sensitivity to adversarial perturbations
    c. Insights on the expressivity with depth

I. Methodology

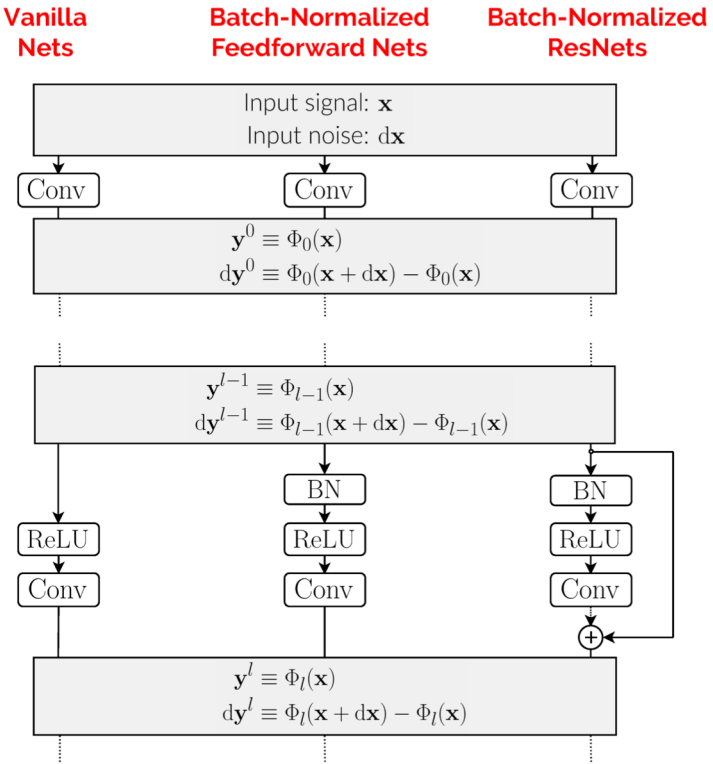
Propagation

Simultaneous propagation of:

- The standard DNN **signal**
- A small **additive noise** corrupting this signal

We will focus on 3 setups :

- **Vanilla Nets:**
  conv + ReLU
- **Batch-Normalized Feedforward Nets:**
  conv + batch norm + ReLU
- **Batch-Normalized ResNets:**
  conv + batch norm + ReLU + skip-connections



Data Randomness (I)

First, we introduce the **source of randomness** coming from the **data**:

- The input signal $\mathbf{x}$
- The input noise $\mathrm{d}\mathbf{x}$
- The spatial position α

At this point, **model parameters are fixed**, i.e. DNNs define **fixed mappings** from inputs to outputs.

Our goal is to define the potential “bad properties” — referred as **“pathologies”** — of such fixed DNN mappings.

Data Randomness (II)

Effective Rank:

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

The effective rank measures the number of directions in which the signal effectively "lives":

- = 1 when it "lives" on a line
- = 2 when it "lives" isotropically on a plane
- etc

Normalized Sensitivity:

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

The normalized sensitivity measures to which extent the DNN amplifies the noise with respect to the signal:

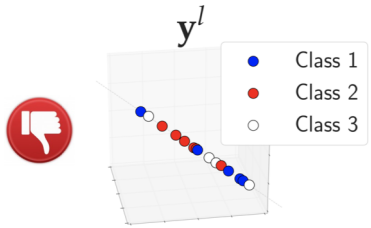
- DNNs with $\chi^l > 1$ are **noise amplifiers**
- DNNs with $\chi^l < 1$ are **noise reducers**

Data Randomness (III)

Pathology of One-Dimensional Signal:

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

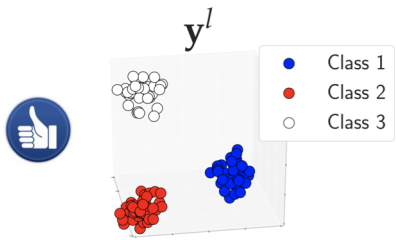
- The signal becomes **line-like concentrated**
- This pathology is incompatible e.g. with the one-hot target of multiclass classification (with effective rank typically equal to the number of classes minus 1)

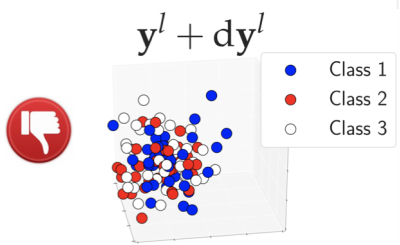


Pathology of Exploding Sensitivity:

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

- The signal becomes **drowned** in the noise
- This pathology might be compatible with low training loss but any corruption $\mathrm{d}\mathbf{x}$ on the test set will lead to the corrupted signal $\mathbf{y}^l + \mathrm{d}\mathbf{y}^l$ becoming pure noise





Model Parameters Randomness (I)

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

This initialization is supposed **standard**:

- Weights initialized following He et al. (2015)
- Biases initialized with 0's
- Scale batch norm parameters initialized with 1's
- Shift batch norm parameters initialized with 0's

A high probability of pathology at initialization likely **prevents training from proceeding normally.**

Model Parameters Randomness (II)

Most results still hold with an initialization **uniform in the constrained parameter space**:

- With channels sampled independently
- With weights, biases, scale and shift batch norm parameters sampled uniformly in hyperballs

A high probability of pathology under such initialization likely leads to:

- **Untrainability** if pathologies are incompatible with low training loss (e.g. one-dimensional signal)
- **Poor generalization** if pathologies are compatible with low training loss but unlikely to generalize (e.g. exploding sensitivity)

Further Notation and Assumption (I)

We denote geometric increments between layer l-1 and l with the prefix δ, i.e. $\delta\chi^l = \chi^l / \chi^{l-1}$.

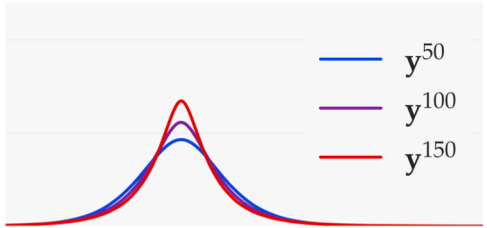
We write $a \lesssim b$ when $a(1+\varepsilon_a) \leq b(1+\varepsilon_b)$ with $|\varepsilon_a| \ll 1$, $|\varepsilon_b| \ll 1$ with high probability.

Further Notation and Assumption (II)

We assume that the **width** (i.e. the number of channels) **is large**:

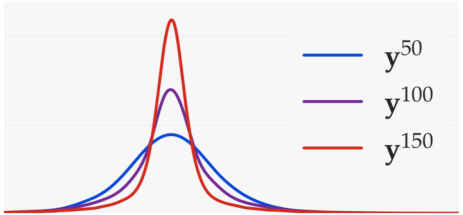
- This assumption always holds in practice
- In contrast, the "mean field" approximation of DNNs as Gaussian Processes breaks down for feedforward nets at high depth

Vanilla Nets



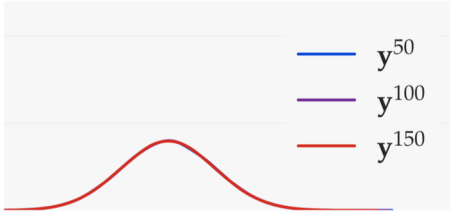
CNNs of width 512
(fixed data, random model params)

Batch-Normalized Feedforward Nets



CNNs of width 512
(fixed data, random model params)

Batch-Normalized ResNets



CNNs of width 512
(fixed data, random model params)

II. Results

Vanilla Nets

Theory

Geometric increments $\delta\chi^l$ are **bounded**:

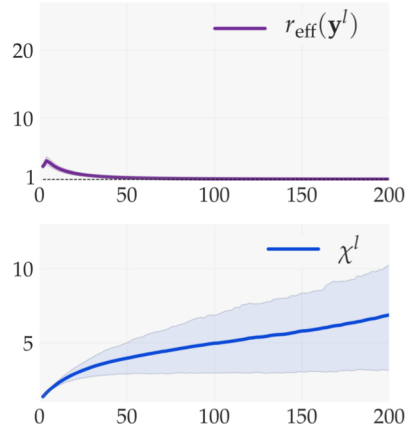
$$1 \lesssim \delta\chi^l \lesssim \sqrt{2}$$

Only two possibilities of evolution:

1. Pathology of exploding sensitivity
2. Concentration of the signal on the **semi-line** generated by its average vector:
   a. Pathology of one-dimensional signal
   b. Subexponential growth of the normalized sensitivity χ^l
   c. Pseudo-linearity, with each additional layer l arbitrarily well approximated by a linear mapping

Experiments

We observe the possibility 2.



Batch-Normalized Feedforward Nets

Theory

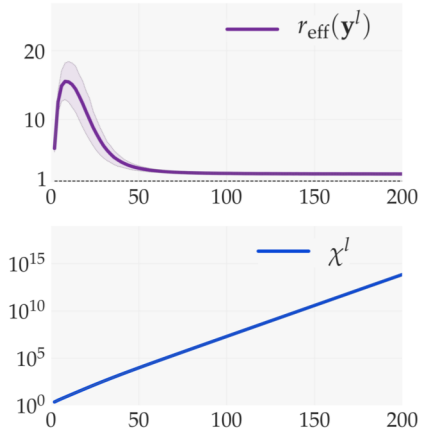
- Geometric increments $\delta\chi^l$ are **bounded**:

$$\delta\chi^l \equiv \delta_{\mathrm{BN}}\chi^l \cdot \delta_\phi\chi^l \;\text{ with }\; 1 \lesssim \delta_{\mathrm{BN}}\chi^l \lesssim C \text{ and } 1 \lesssim \delta_\phi\chi^l \lesssim \sqrt{2}$$

- If the signal has few directions of variance, then $|\delta\chi^l{}_{\mathrm{BN}}\text{-}1|$ is non-negligible
- If the signal is not too fat-tailed, then $|\delta\chi^l{}_{\varphi}\text{-}1|$ is non-negligible

Experiments

- Only few directions of signal variance at high depth
- Pathology of exploding sensitivity



Batch-Normalized ResNets

Theory

- χ^l evolves as a **power-law**, due to the "dilution" of the residual path into the skip-connection path, with ratio of signal variance $\propto 1/(l{+}1)$:

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

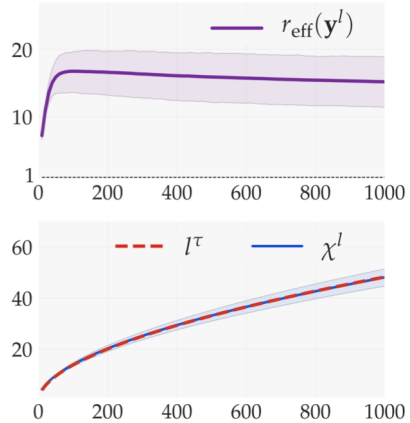
$$C_{\min} l^{\tau_{\min}} \lesssim \chi^l \lesssim C_{\max} l^{\tau_{\max}}$$

- The bounds on χ^l are obtained by integrating the log of the bounds on $\delta\chi^l$
- Batch-normalized resnets are **logarithmic versions** of batch-normalized feedforward nets:

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

Experiments

- Many directions of signal variance
- Perfect power-law fit of χ^l
- Batch-normalized resnets are logarithmic versions of batch-normalized feedforward nets



Other Initialization / Activation

Replacing the standard initialization by the **uniform initialization**:

- For vanilla nets, either pathology of zero-dimensional signal or pathology of one-dimensional signal
- For batch-normalized feedforward nets and batch-normalized resnets, similar results

Replacing **ReLU** by **tanh**:

- For vanilla nets, either pathology of zero-dimensional signal (“ordered phase”) or pathology of exploding sensitivity (“chaotic phase”)
- For batch-normalized feedforward nets and batch-normalized resnets, similar results

All cases are equivalent in terms of presence / absence of pathologies.

Other Architectures

Other normalized feedforward nets

- Instance Normalization (Ulyanov et al., 2016): similar to Batch Normalization
- Layer Normalization (Ba et al., 2016): zero-dimensional signal
- Group Normalization (Wu & He, 2017): in between Layer Normalization and Instance Normalization

Normalized resnets are always logarithmic versions of **normalized feedforward nets.**

Unnormalized resnets

- Exploding activations
- One-dimensional signal

III.Conclusions

Suitable Choices of Architectures

The main force attracting towards pathology is the multiplicativity of feedforward layer composition (conv and ReLU layers can be seen respectively as multiplication by a random matrix and multiplication by a random Bernouilli vector)

- Vanilla nets and batch-normalized feedforward nets are **pathological at high depth** since they are subject to plain feedforward multiplicativity
- Batch-normalized resnets remain **well-behaved at all depths** since the decaying ratio of signal variance $\propto 1/(l+1)$ between residual and skip-connection paths effectively counters feedforward multiplicativity

Our analysis provides theoretical backing to the choices of hyperparameters leading to state-of-the-art performance.

Adversarial Vulnerability

In all setups of ReLU nets, we find $\chi^l \gtrsim 1$:

- An overwhelming part of the hypothesis space has sensitivity larger than the typical variation in output
- Equivalently, an overwhelming part of the hypothesis space is more sensitive to adversarial perturbations than identity mappings
  - ➢ Identity mappings are themselves highly sensitive to adversarial perturbations in high dimension
  - ➢ e.g. setting the 1^{st} dimension equal to 0 in input — a very small perturbation in high dimension — leads to the 1^{st} dimension equal to 0 in output
- So an overwhelming part of the hypothesis space is **highly sensitive to adversarial perturbations** in high dimension

Expressivity

There is a **gap of expressivity** between vanilla nets and batch-normalized feedforward nets:

- For vanilla nets, an overwhelming part of the hypothesis space has subexponential expressivity with depth
- For batch-normalized feedforward nets, an overwhelming part of the hypothesis space has exponential expressivity with depth
- This **gap is not absolute** since any batch-normalized feedforward net can be expressed as a vanilla net by merging batch norm into conv layers
  - ➢ So, there is no contradiction with early works on the exponential advantage of depth over width for vanilla nets (Telgarsky, 2015; Telgarsky, 2016; Bianchini et al. ,2014; Raghu et al., 2017; Poole et al., 2016; Bianchini et al., 2014, Montufar et al., 2014)
- **This gap is still of significance** since it is likely observed at initialization and after training

Thank you!

Paper: https://arxiv.org/abs/1811.03087

Blog Post: https://towardsdatascience.com/its-necessary-to-combine-batch-norm-and-skip-connections-e92210ca04da

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