UNIVERSITEIT STELLENBOSCH UNIVERSITY

SAMSA

Initialising deep noisy ReLU networks

CSIR
our future through science

Arnu Pretorius

Elan van Biljon

Steve Kroon

Herman Kamper

mih media lab.

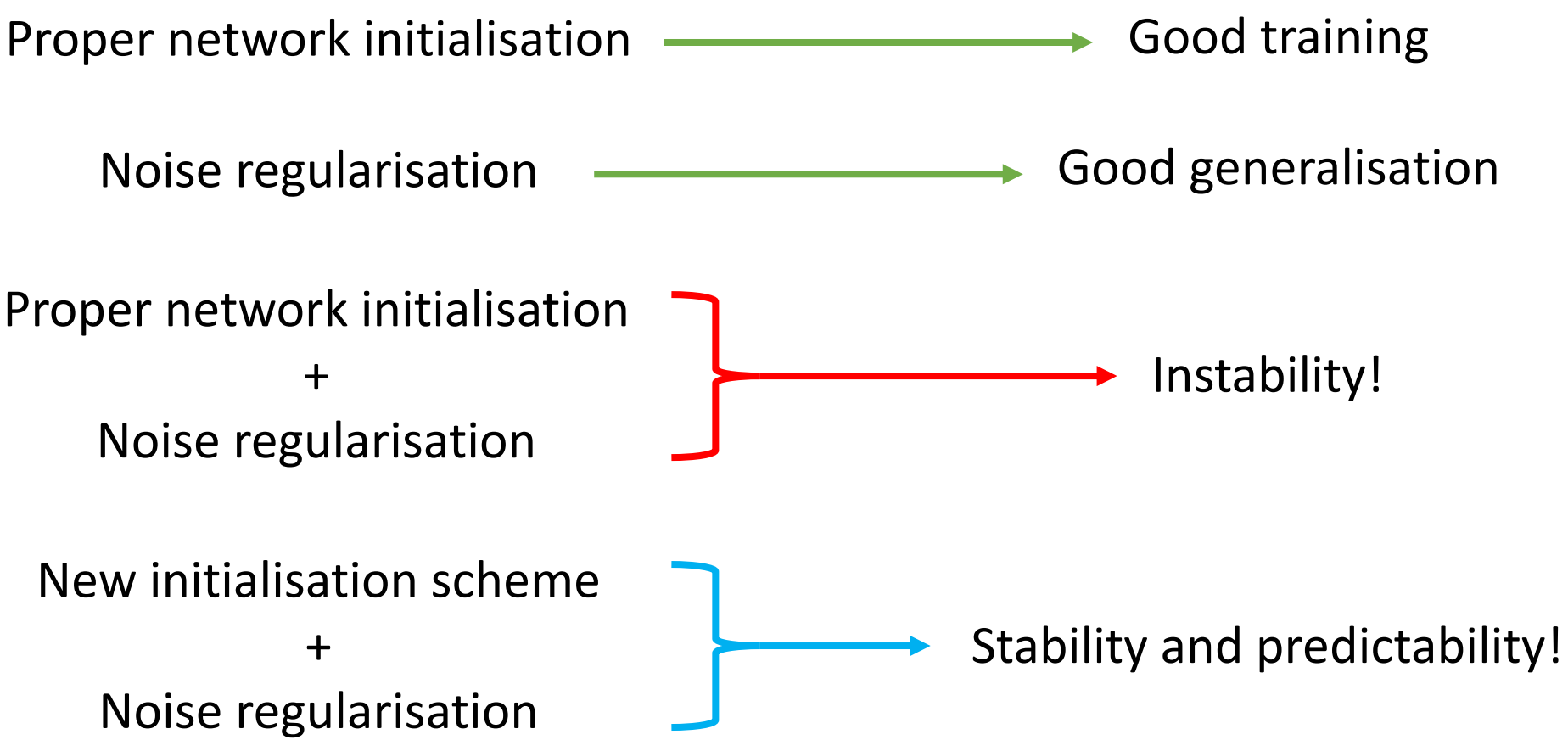
NIPS

Neural Information Processing Systen

cair
CENTRE FOR ARTIFICIAL INTELLIGENCE RESEARCH

Based on the paper: “Critical initialisation for deep signal propagation in noisy rectifier neural networks”, which is to be presented at NIPS 2018.

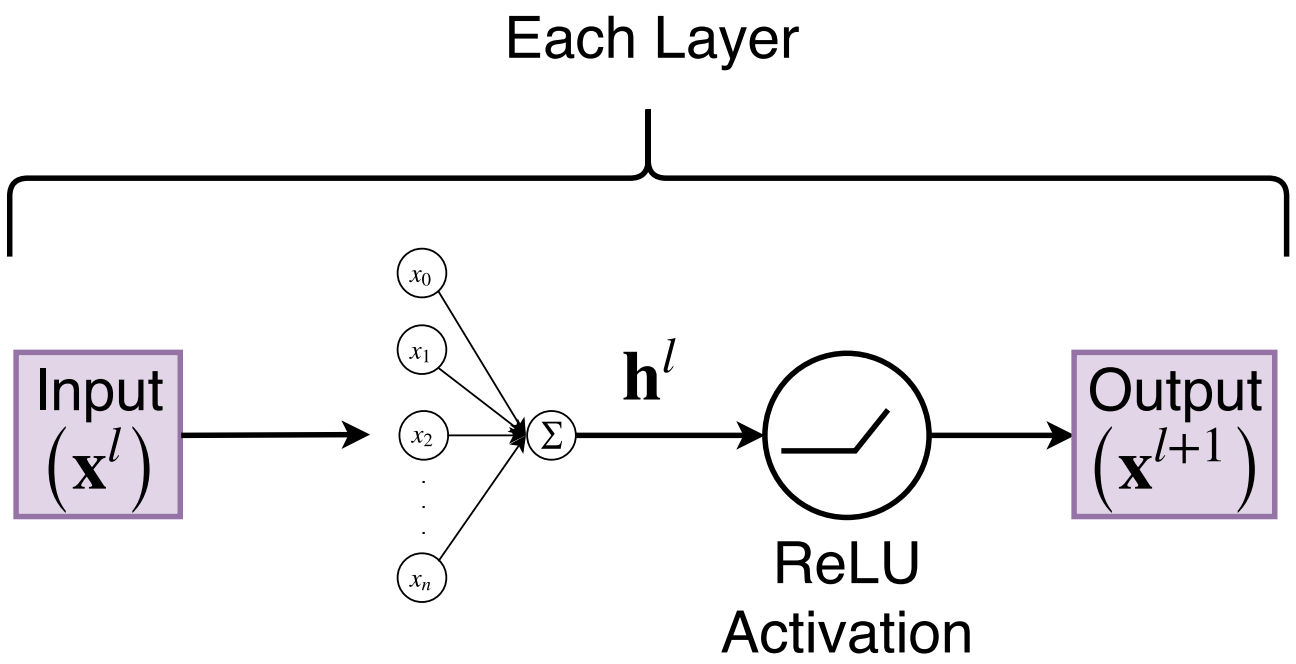
Quick overview



Feed-forward Networks

$$\mathbf{h}^l = W^l \mathbf{x}^l + \mathbf{b}^l$$

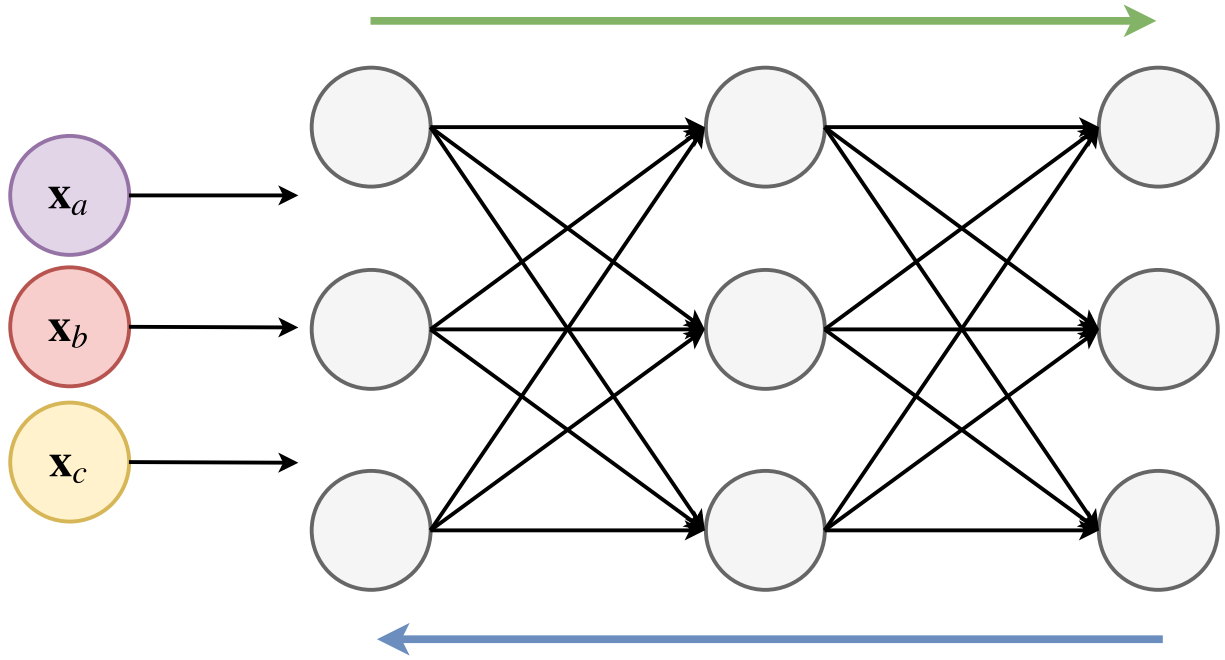
$$\mathbf{x}^{l+1} = \phi\left(\mathbf{h}^l\right) = \mathrm{ReLU}\left(\mathbf{h}^l\right) = \max\left(0, \mathbf{h}^l\right)$$



Necessary conditions for networks to train

"information about the inputs should be able to propagate forward (1)
through the network, and information about the gradients should be (2)
able to propagate backwards through the network."
Schoenholz et al. (2017)

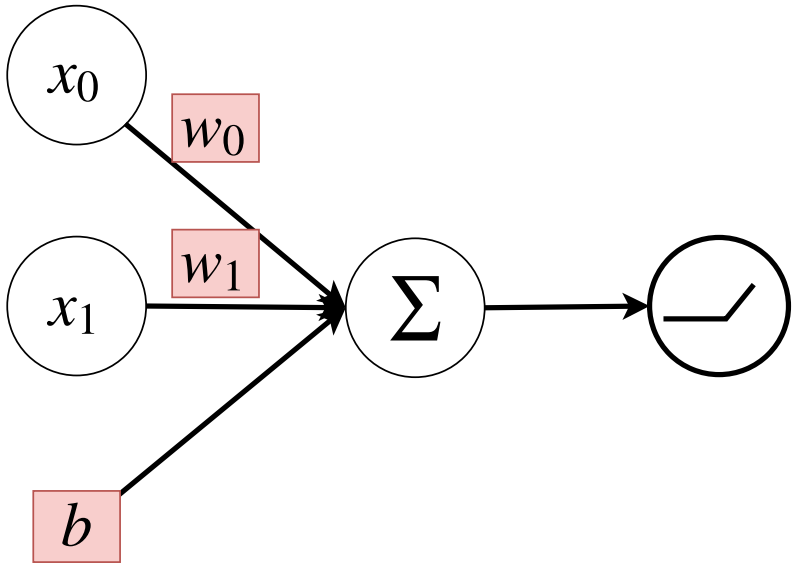
I focus on the first condition



S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, "Deep information propagation," Proceedings of the International Conference on Learning Representations , 2017.

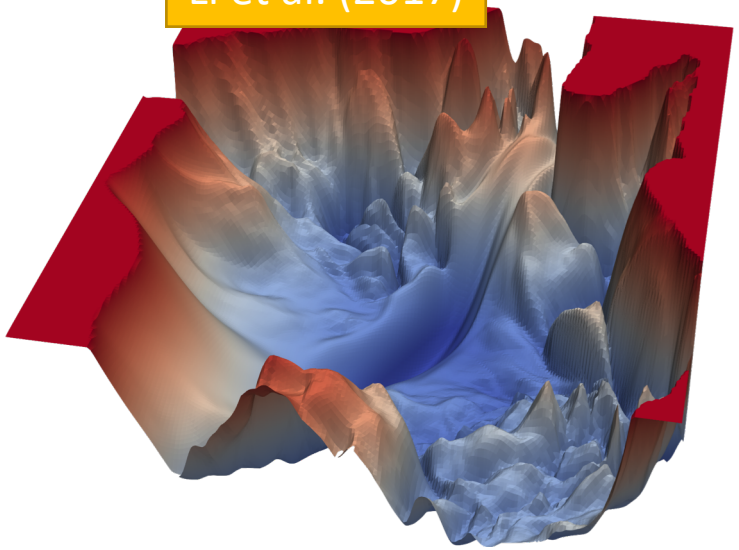
Network initialisation

What?
How we chose our initial weights and biases



Why?
1) Affects starting position on the loss landscape
2) Information flow properties
   1) Input information reaches output
   2) Gradients can propagate back to input

Li et al. (2017)



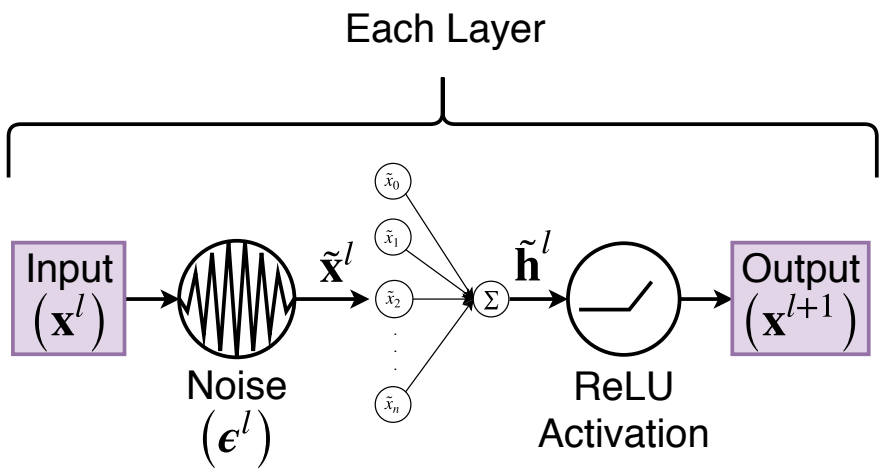
Trained networks of up to 10 000 layers!
Xiao et al. (2018)

L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, “Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks,” Proceedings of the International Conference on Machine Learning , 2018.

H. Li, Z. Xu, G. Taylor, and T. Goldstein, “Visualizing the loss landscape of neural nets”, arXiv preprint arXiv:1712.09913, 2017.

Noise regularisation

What?
Corrupting the input to each layer with some noise

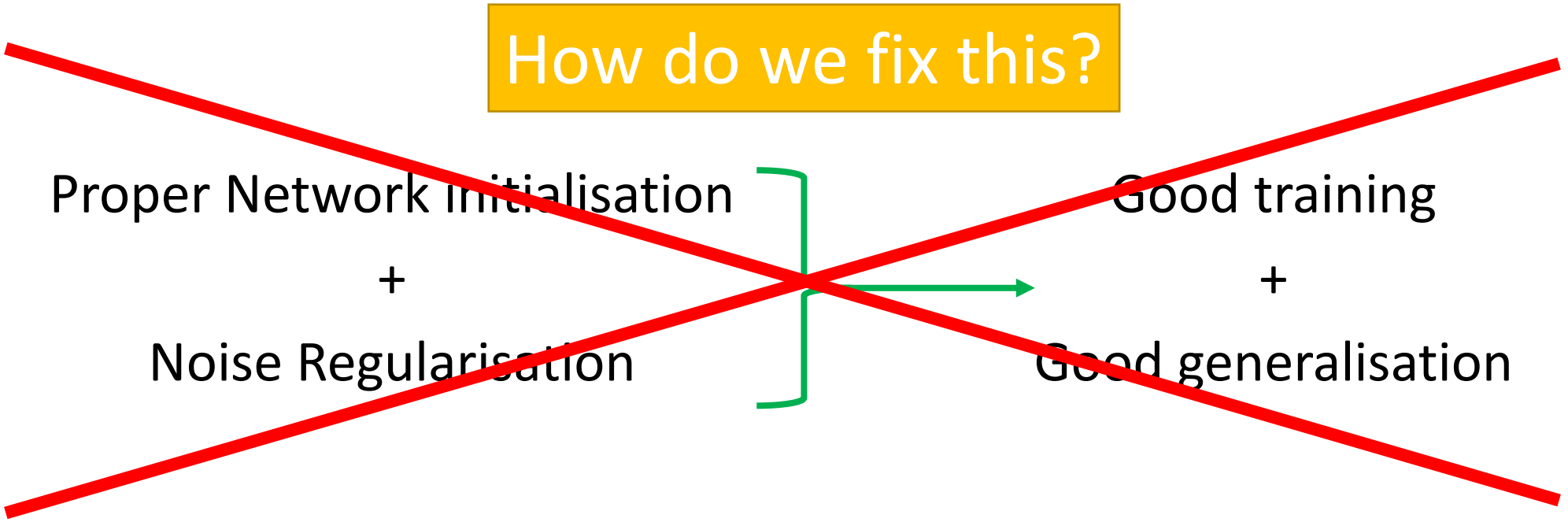


$\mathbf{x}^l$ $\quad$ $\boldsymbol{\epsilon}^l$ $(p = 0.6)$ $\quad$ $\tilde{\mathbf{x}}^l$

3		1		3
1		0		0
4	×	1	=	4
1		0		0
5		1		5
9		1		9

Why?
Improves generalisation (performance on unseen data)
- Avoids overfitting

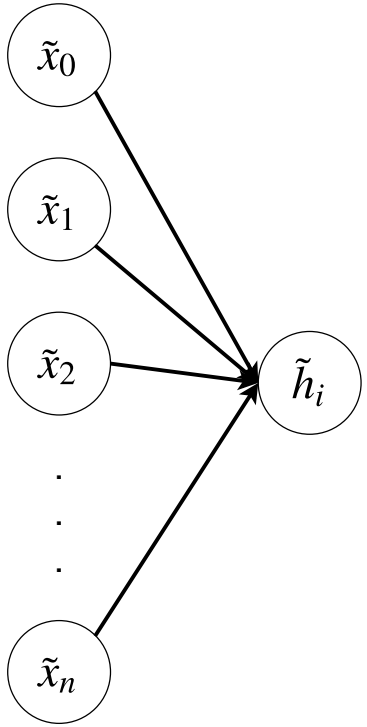
What is the problem?



Mean field theory

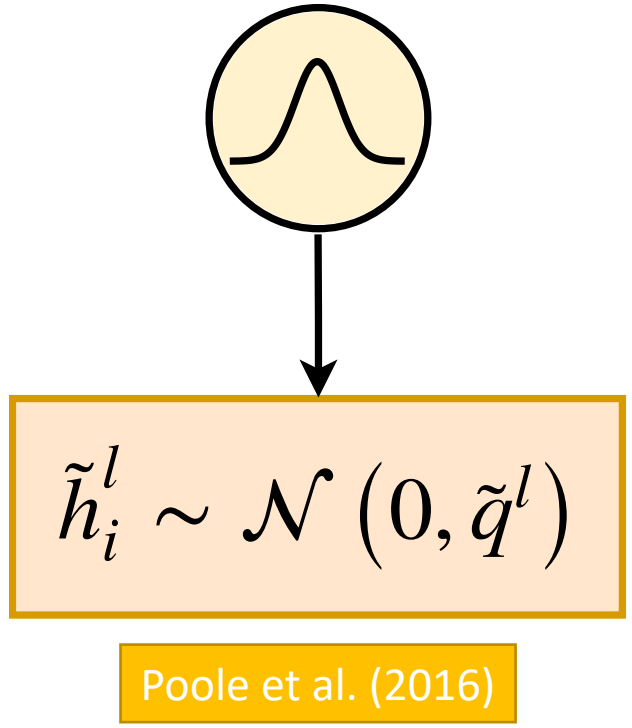
Mean field assumption: wide network layers

- Sum of many independent RVs
- Central limit theorem
- Pre-activations are normally distributed mean zero



Look at "average" / expected behavior

- We look at the expected variance



Poole et al. (2016)

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, "Exponential expressivity in deep neural networks through transient chaos," in Advances in Neural Information Processing Systems, 2016, pp. 3360–3368.

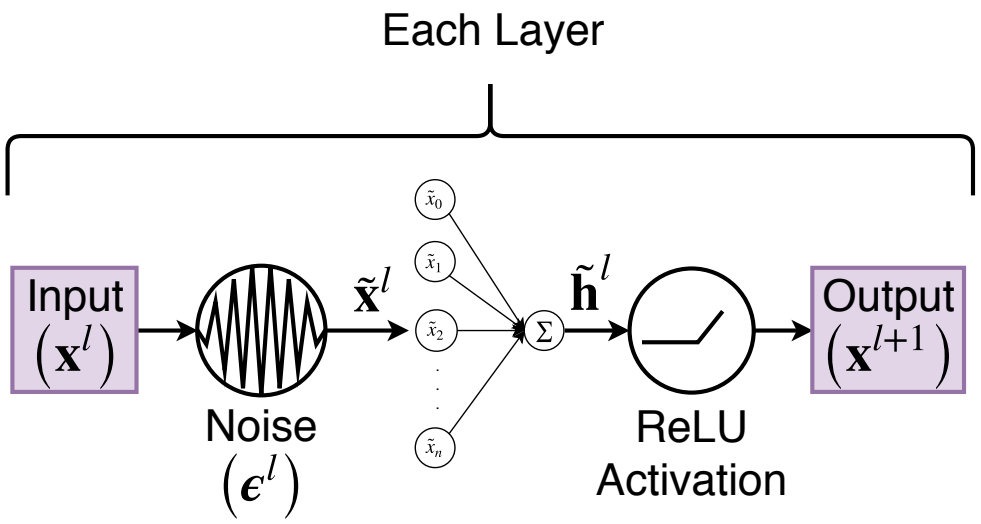
Network set up

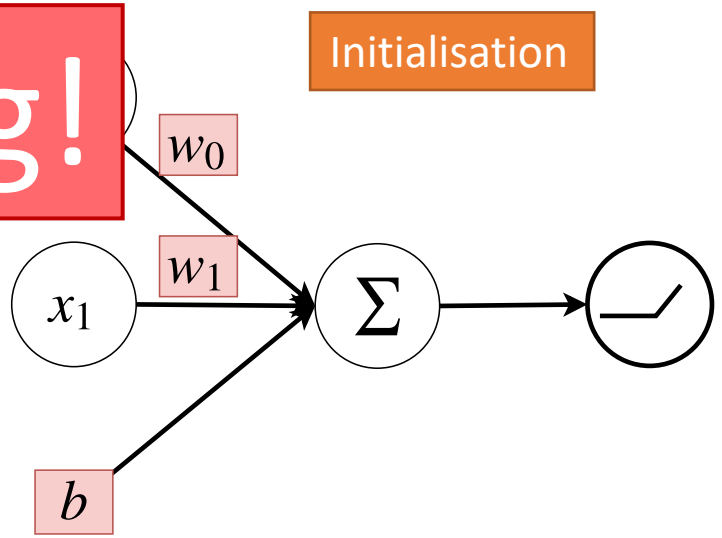
Layer topology

No training!

Initialisation

$$\tilde{\mathbf{h}}^l = W^l \left(\mathbf{x}^l \odot \boldsymbol{\epsilon}^l \right) + \mathbf{b}^l$$





$$W_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_w^2}{D_{l-1}}\right)$$

$$b_i \sim \mathcal{N}\left(0, \sigma_b^2\right)$$

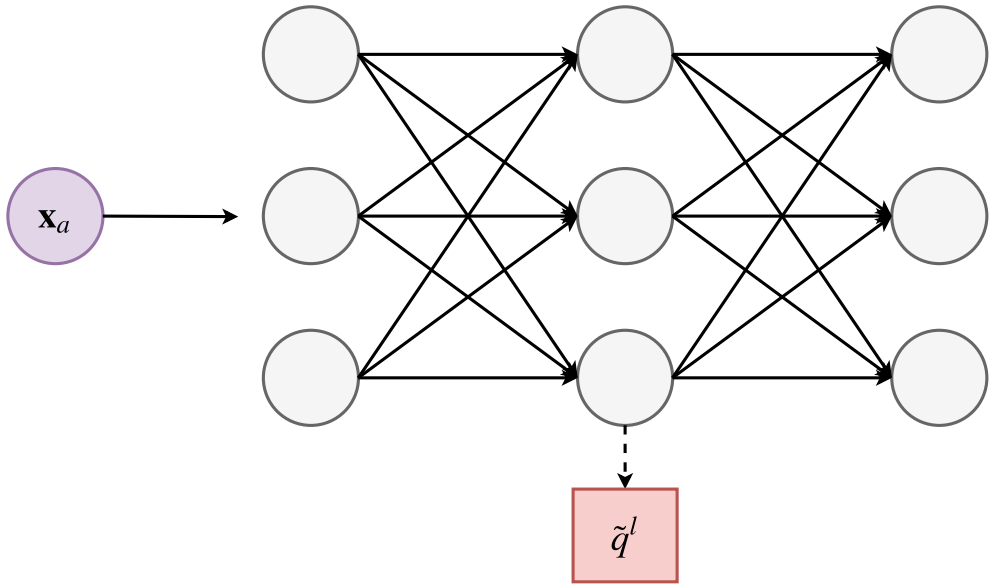
Mean field theory – single input information

$$\tilde{q}^l = \sigma_w^2 \mathrm{E}_z \left[\phi\left(\sqrt{\tilde{q}^{l-1}} z\right)^2 \odot \mu_2 \right] + \sigma_b^2$$

$$\tilde{q}^l = \sigma_w^2 \left[\frac{\tilde{q}^{l-1}}{2} \odot \mu_2 \right] + \sigma_b^2$$

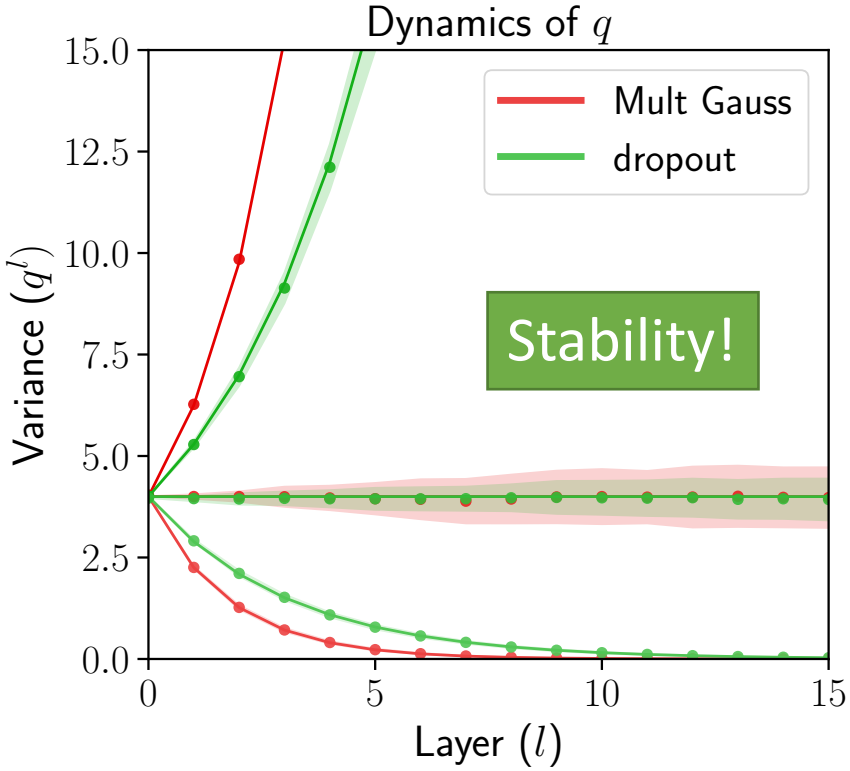
$$\tilde{q}^l = \tilde{q}^{l+1} = \tilde{q}^*$$

$$(\sigma_w^2, \sigma_b^2) = \left(\frac{2}{\mu_2}, 0\right)$$

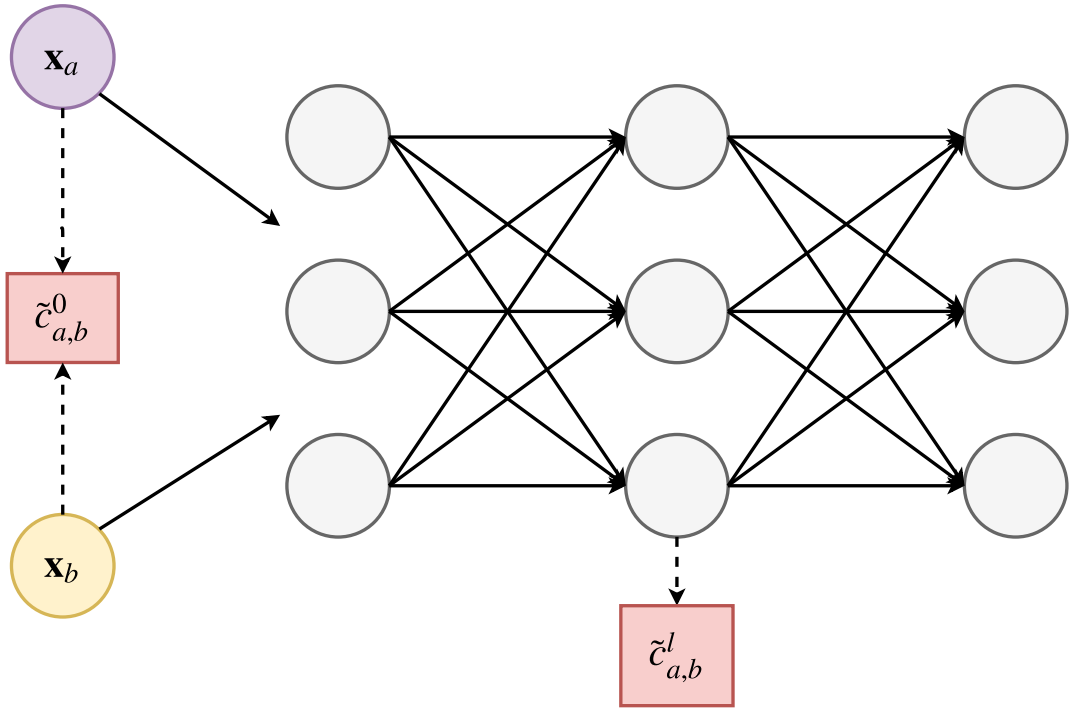


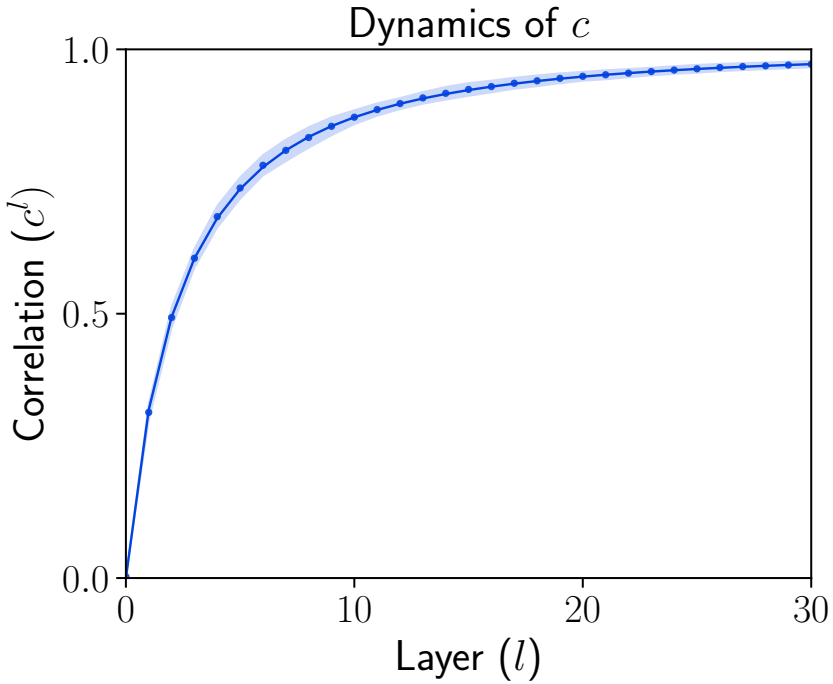
Part of the first condition is met!

inputs should be able to propagate forward



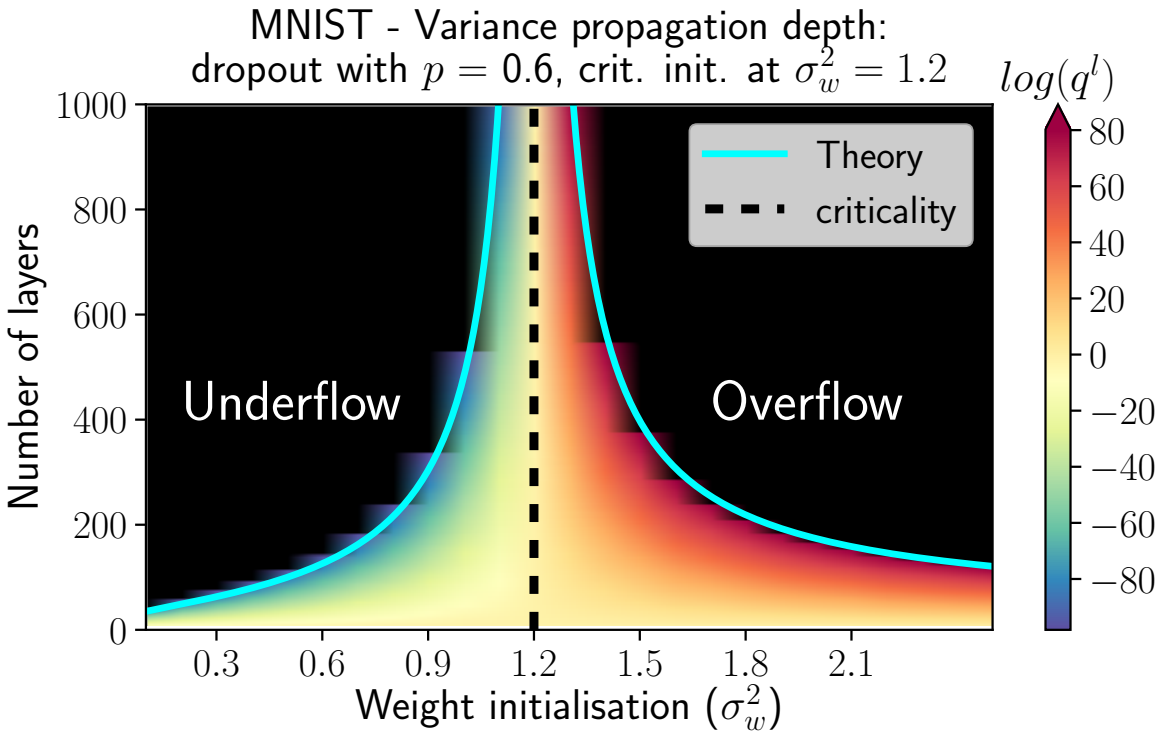
Mean field theory – dual input information

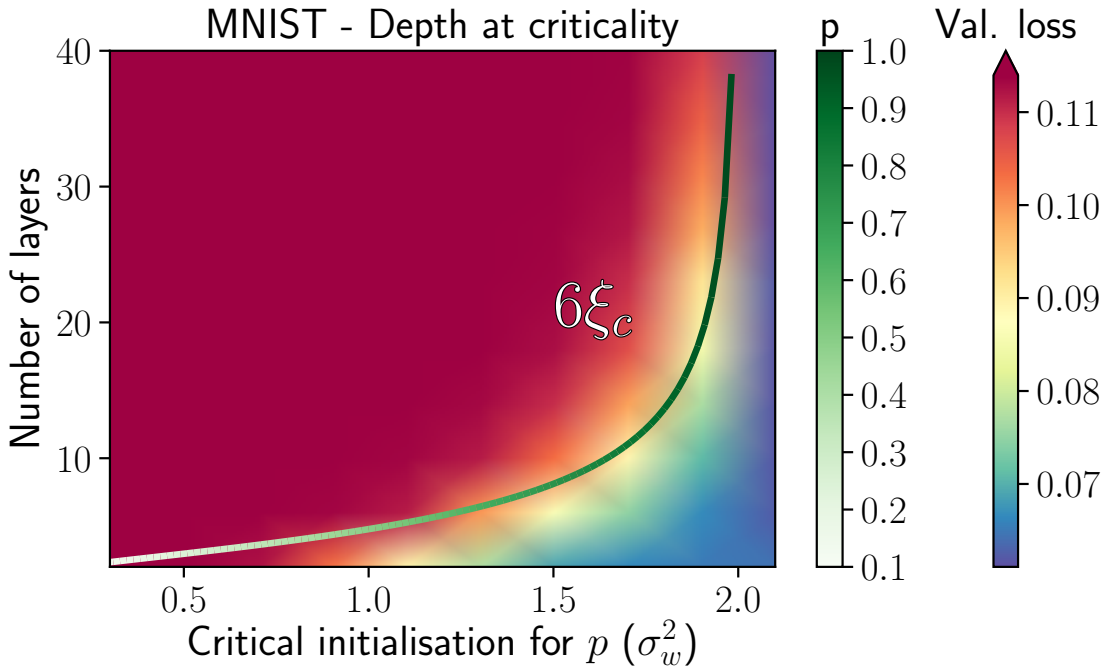




Lose discriminatory information!
First condition is not fully met :(

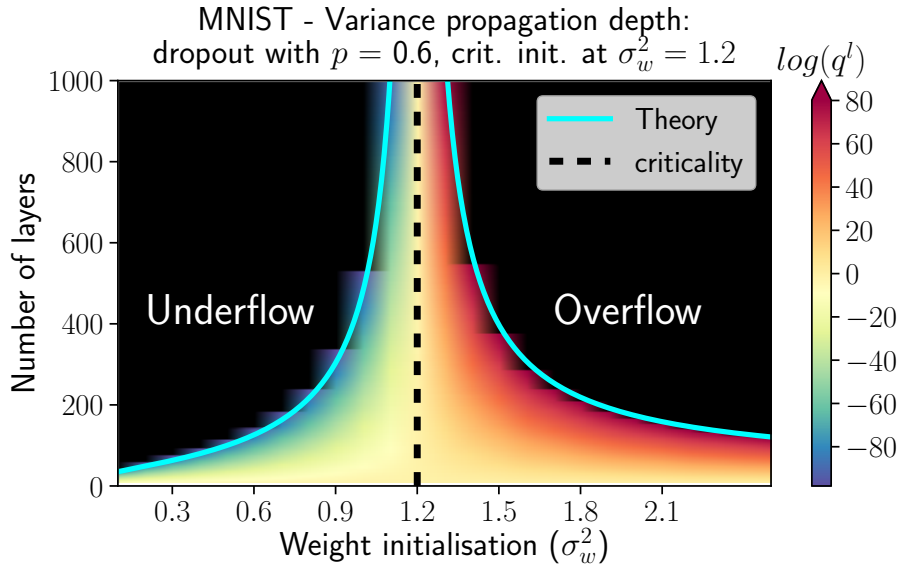
Real-world experiments

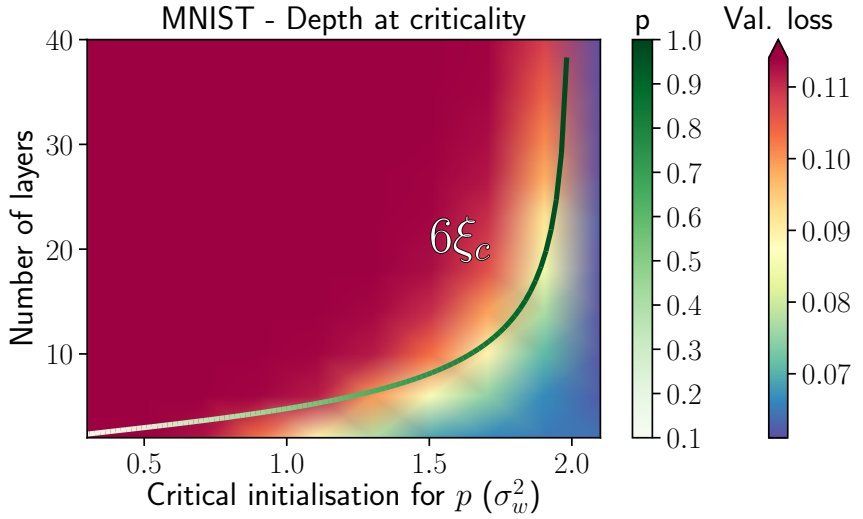




$$\text{Dropout} \rightarrow \left(\sigma_w^2, \sigma_b^2\right) = (2p, 0)$$

Special cases





$$\text{Dropout} \rightarrow \left(\sigma_w^2, \sigma_b^2\right) = (2p, 0)$$

$$p = 0.5 \Rightarrow \sigma_w^2 = 1 \rightarrow \text{Xavier}$$

Simonyan and Zisserman (2014); Glorot and Bengio (2010)

$$p = 1 \Rightarrow \sigma_w^2 = 2 \rightarrow \text{He}$$

He et al. (2015)

Why did p=0.5 work so well in the past?

$$\text{Dropout} \rightarrow \left(\sigma_w^2, \sigma_b^2\right) = (2p, 0)$$

$$p = 0.5 \Rightarrow \sigma_w^2 = 1 \rightarrow \text{Xavier}$$

Simonyan and Zisserman (2014); Glorot and Bengio (2010)

$$p = 1 \Rightarrow \sigma_w^2 = 2 \rightarrow \text{He}$$

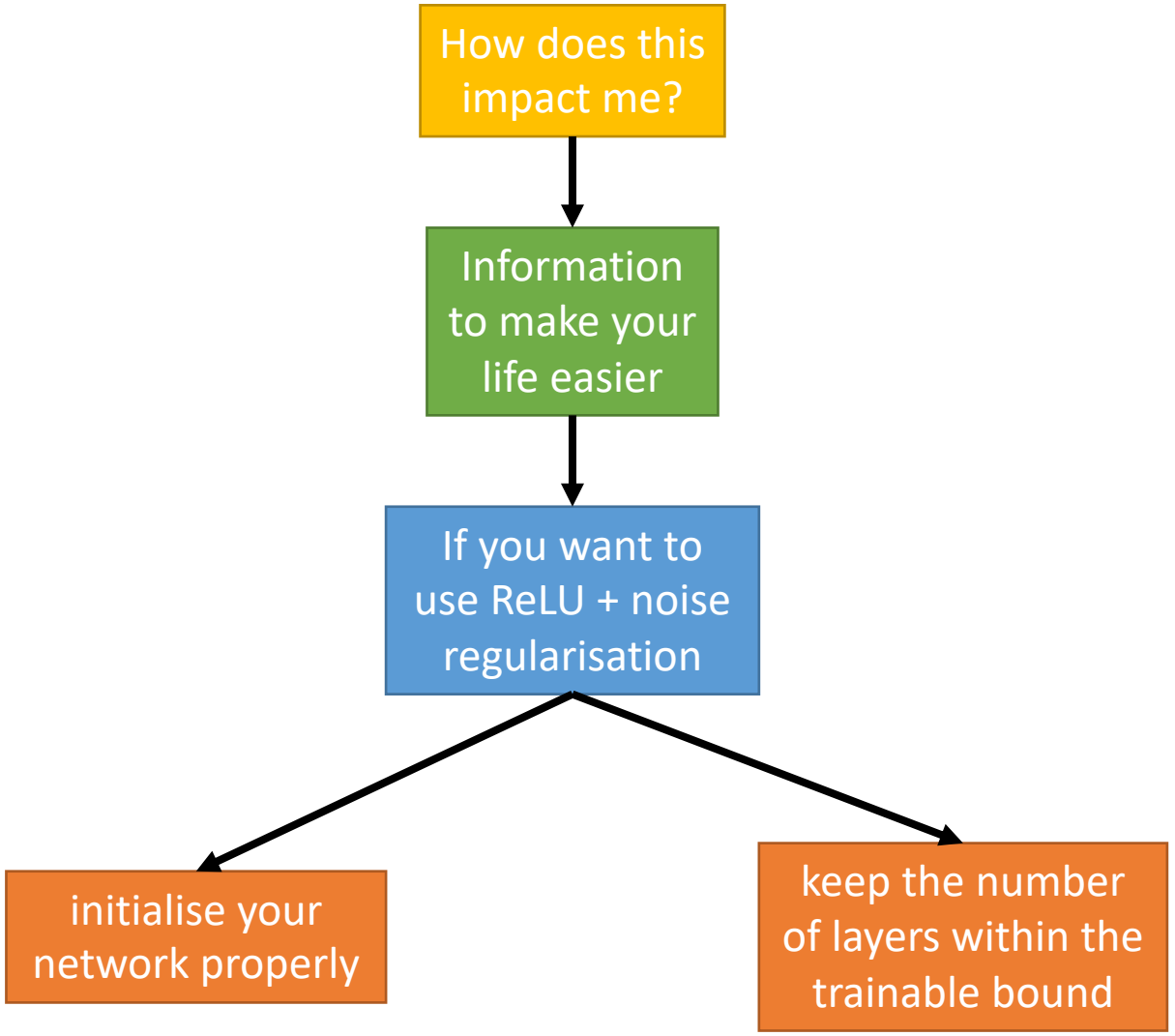
He et al. (2015)

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556 , 2014.

X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the International Conference on Artificial Intelligence and Statistics , 2010, pp. 249–256.

K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proceedings of the IEEE International Conference on Computer Vision , 2015, pp. 1026–1034.

Conclusions



References

Our Paper:

arXiv:

1811.00293

GitHub:

noisy_signal_prop

Acknowledgements:
We would like to thank Google, the CSIR/SU Centre for Artificial Intelligence Research (CAIR) and we gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

Bibliography:

S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein, “Deep information propagation,” Proceedings of the International Conference on Learning Representations , 2017.

L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. S. Schoenholz, and J. Pennington, “Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks,” Proceedings of the International Conference on Machine Learning , 2018.

H. Li, Z. Xu, G. Taylor, and T. Goldstein, “Visualizing the loss landscape of neural nets”, arXiv preprint arXiv:1712.09913, 2017.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli, “Exponential expressivity in deep neural networks through transient chaos,” in Advances in Neural Information Processing Systems, 2016, pp. 3360–3368.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv preprint arXiv:1409.1556 , 2014.

X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in Proceedings of the International Conference on Artificial Intelligence and Statistics , 2010, pp. 249–256.

K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification,” in Proceedings of the IEEE International Conference on Computer Vision , 2015, pp. 1026–1034.

Thank you!

Questions?