Critical initialisation for deep signal propagation in noisy rectifier neural networks

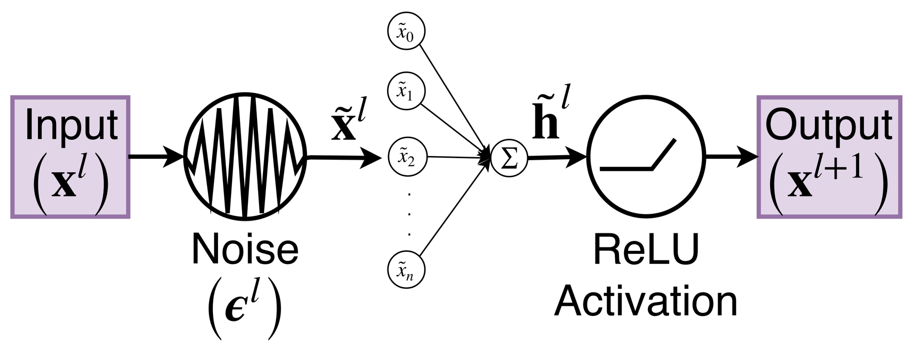
Arnu Pretorius*‡ Elan Van Biljon* Steve Kroon* Herman Kamper†

*Computer Science, †E&E Engineering, Stellenbosch University; ‡CSIR/SU Centre for Artificial Intelligence Research, South Africa

Contributions

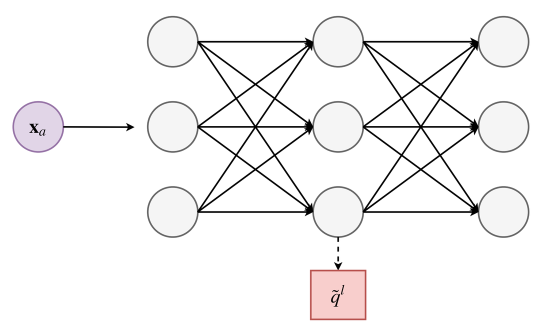
- We extend the mean field framework developed by Poole et al. (2016) and Schoenholz et al. (2017), to describe noisy signal propagation in fully connected feed-forward neural networks.
- We derive variance critical weight initialisation strategies for noisy ReLU networks, suitable for a wide range of noise models.
- We describe the limitations to information flow as a result of noise by studying signal correlation dynamics.

1. Noisy signal propagation



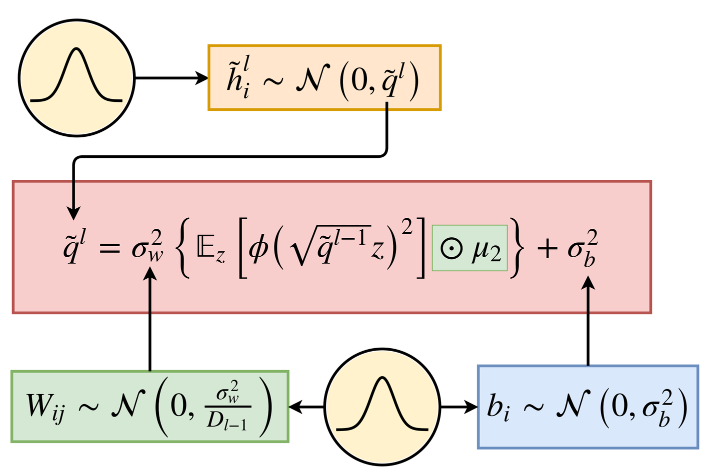
$$\tilde{\mathbf{h}}^l = W^l\left(\mathbf{x}^l \odot \boldsymbol{\epsilon}^l\right) + \mathbf{b}^l$$

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



We study signal propagation of an input $\mathbf{x}^0 \in \mathbb{R}^{D_0}$ to a noise regularised deep ReLU network at initialisation, where the weights $W^l \in \mathbb{R}^{D_l \times D_{l-1}}$ and the biases $\mathbf{b}^l \in \mathbb{R}^{D_l}$ are randomly sampled at each layer $l = 1, ..., L$.

2. Critical initialisation for noisy ReLU networks



Sampling the weights and biases from zero mean Gaussian distributions with pre-specified variances σ_w^2/D_{l-1} and σ_b^2, we use the mean field assumption to approximate the distribution of the pre-activations at each layer l, in the large width limit, by a zero mean Gaussian with variance $\tilde{q}^l$. Therefore, each $\tilde{\mathbf{h}}_i^l \sim \sqrt{\tilde{q}^l}z$, where $z \sim \mathcal{N}(0,1)$. The fixed point of the variance recurrence (**red block**) gives the critical initialisation as $(\sigma_w, \sigma_b) = \left(\sqrt{\frac{2}{\mu_2}}, 0\right)$, where μ_2 is the second moment of the noise distribution.

Initialisations for noisy ReLU

NOISE	P(ϵ)	CRITICAL INIT
GAUSSIAN	$\mathcal{N}(1, \sigma_\epsilon^2)$	$(\sigma_w, \sigma_b, \sigma_\epsilon) = \left(\sqrt{\frac{2}{\sigma_\epsilon^2+1}}, 0, \sigma_\epsilon\right)$
LAPLACE	$Lap(1, \beta)$	$(\sigma_w, \sigma_b, \beta) = \left(\sqrt{\frac{2}{2\beta^2+1}}, 0, \beta\right)$
POISSON	$Poi(1)$	$(\sigma_w, \sigma_b, \lambda) = (1, 0, 1)$
DROPOUT	$P(\epsilon = \frac{1}{p}) = p$, $P(\epsilon = 0) = 1 - p$	$(\sigma_w, \sigma_b, p) = (\sqrt{2p}, 0, p)$

3. Initialisations for different noise types

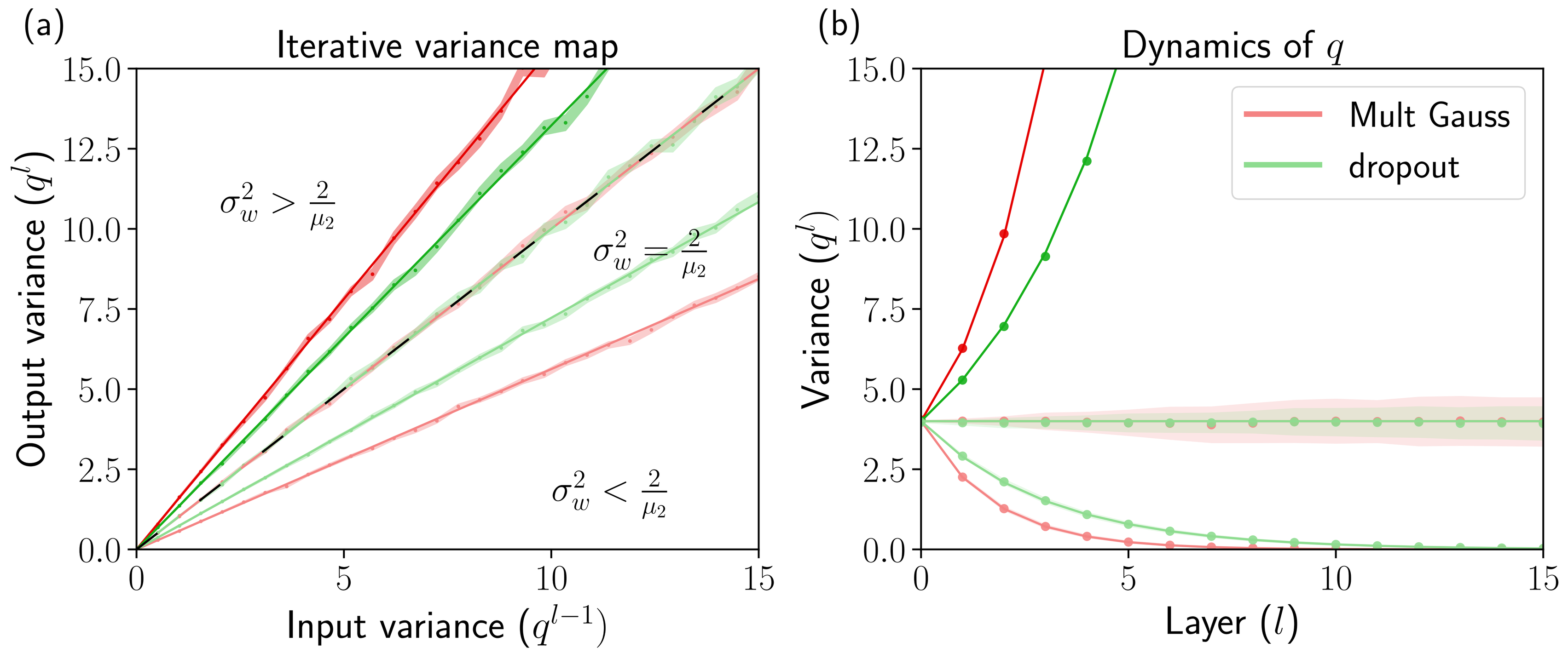


Figure 1: Variance propagation for noisy ReLU.

For schemes not initialising at criticality, the variance map in Figure 1 (a) lies off the identity line and the variances in (b) either explode, or vanish. The critical initialisation lies on the boundary between these two extremes, (as shown in Figure 2) and preserves the signal in (b) throughout the forward pass with roughly constant variance.

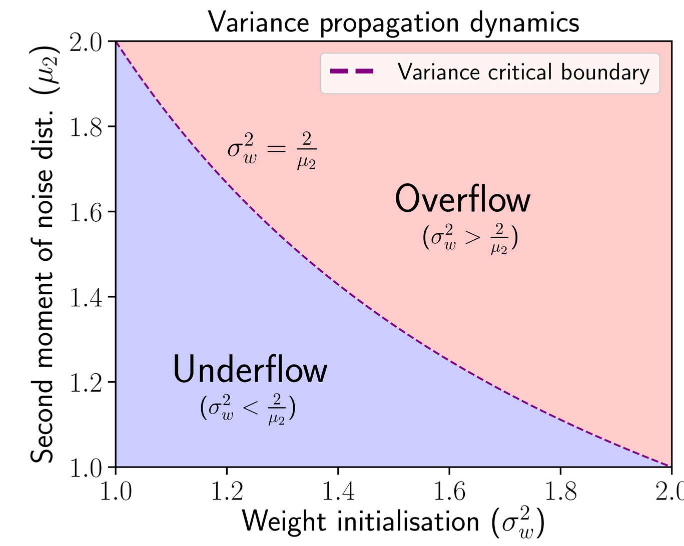
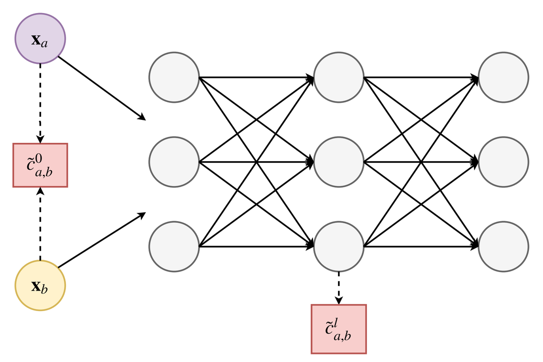


Figure 2: Variance critical boundary.

4. Correlation dynamics



$$\tilde{c}_{a,b}^l = \frac{1}{\pi\mu_2}\left\{\tilde{c}_{a,b}^{l-1}\sin^{-1}\left(\tilde{c}_{a,b}^{l-1}\right) + \sqrt{1-\left(\tilde{c}_{a,b}^{l-1}\right)^2} + \frac{\pi}{2}\tilde{c}_{a,b}^{l-1}\right\}$$

We further study the dynamics of correlation $\tilde{c}_{ab}^l$ between two inputs $\mathbf{x}_a$ and $\mathbf{x}_b$. At large depth, inputs end up uniformly correlated irrespective of their starting correlation, as shown in Figure 3 (a) and (b). Therefore, random deep ReLU networks lose discriminatory information about their inputs as the depth of the network increases.

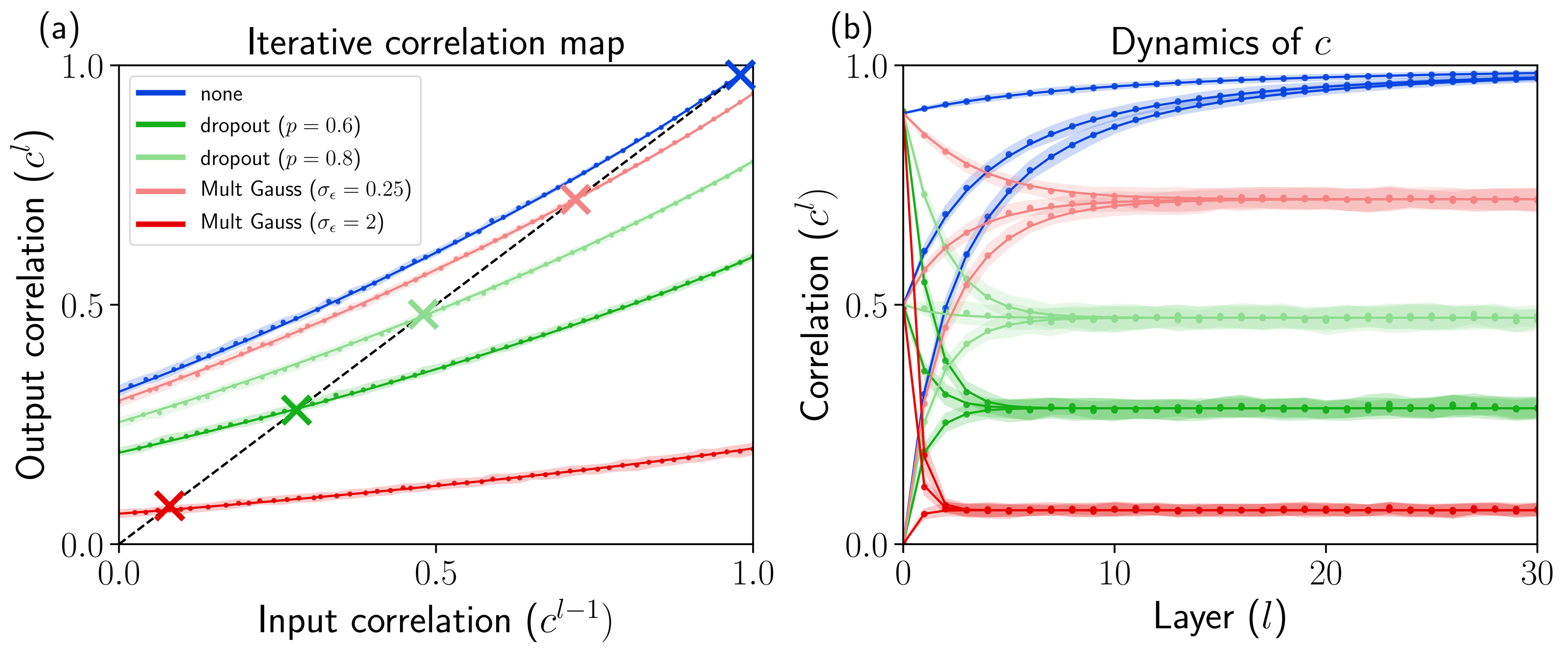


Figure 3: Correlation dynamics for noisy ReLU.

5. Experiments with dropout on real-world data

A random network loses useful information more quickly when injected with noise, therefore we examine trainable depths on MNIST using dropout.

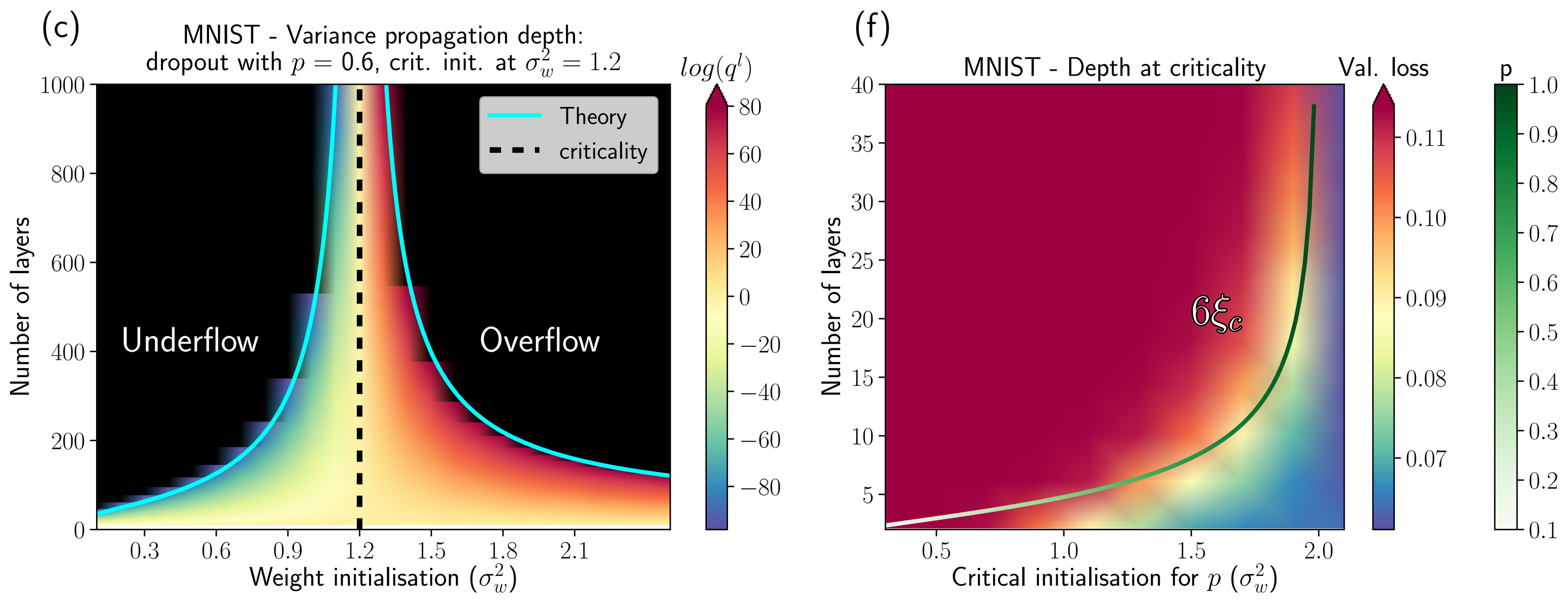


Figure 4: Depth scale experiments on MNIST.

Please scan the following QR-codes for additional results in the paper, source code and our video.

Paper Code Video

Takeaways

- When using multiplicative noise (e.g. dropout) to regularise a deep ReLU network, initialising the weights and biases from normal distributions with standard deviations $(\sigma_w, \sigma_b) = \left(\sqrt{\frac{2}{\mu_2}}, 0\right)$ (where μ_2 is the second moment of the noise distribution), ensures reliable signal propagation.
- However, even at criticality, noise causes the correlation between signals to decay with increasing depth and this limits the depth at which noisy ReLU networks are able to perform well.

Acknowledgements

We would like to thank NeurIPS, Google, the CSIR/SU Centre for Artificial Intelligence Research (CAIR) as well as the Science Faculty and the Postgraduate and International Office of Stellenbosch University for financial support. Finally, we gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research.

References

[1] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. NeurIPS, 2016.

[2] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep Information Propagation. ICLR, 2017.

Presented at the 32nd Conference on Neural Information Processing Systems (NeurIPS), 2018

Contact: arnupretorius@gmail.com, elanvanbiljon@gmail.com