Deep signal propagation for noisy rectifier neural networks

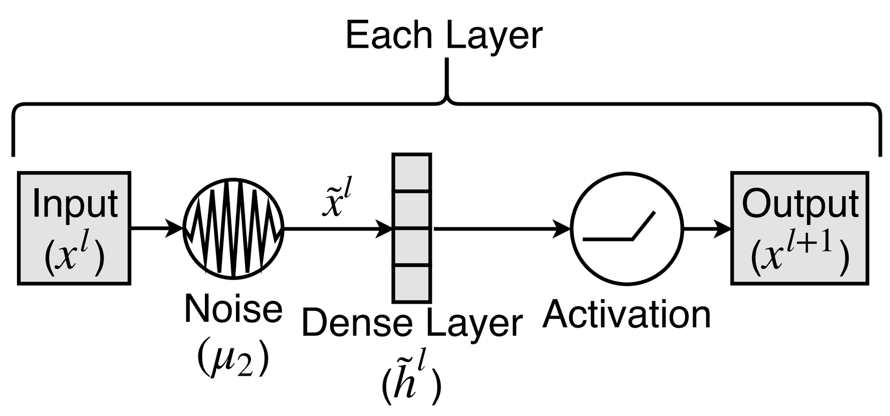
Arnu Pretorius Elan Van Biljon Steve Kroon Herman Kamper*

Division of Computer Science and EE Engineering*, Stellenbosch University

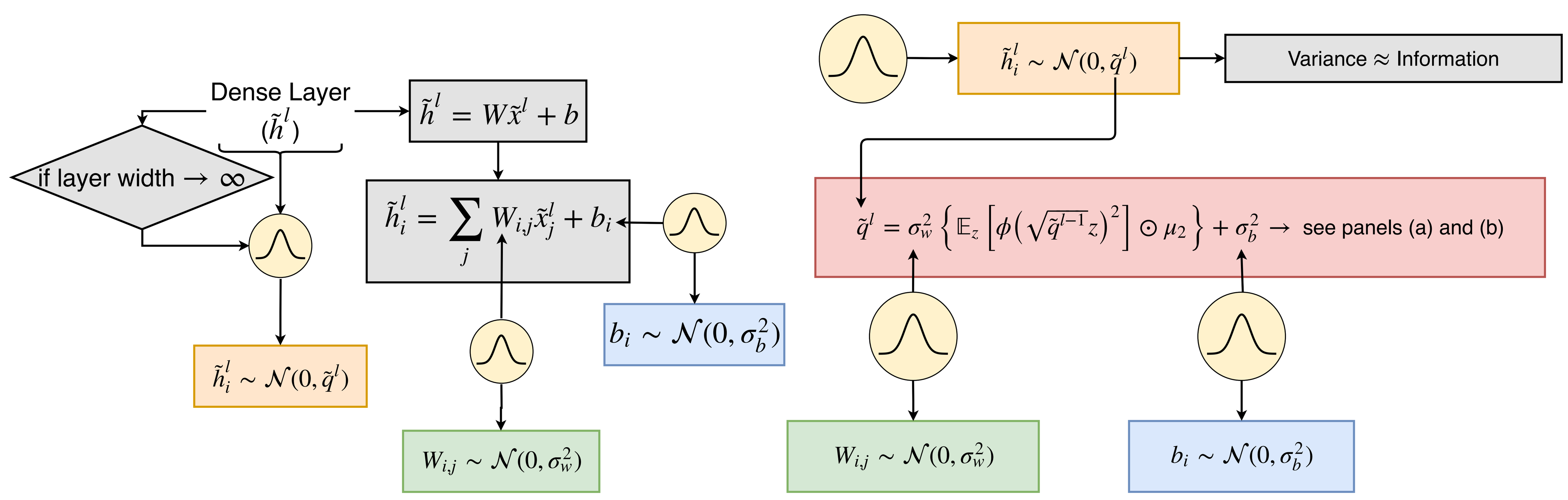
Contributions

- We extend the mean field framework in [1], 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 the signal correlation dynamics.

Noisy signal propagation model



Critical initialisation for noisy ReLU networks

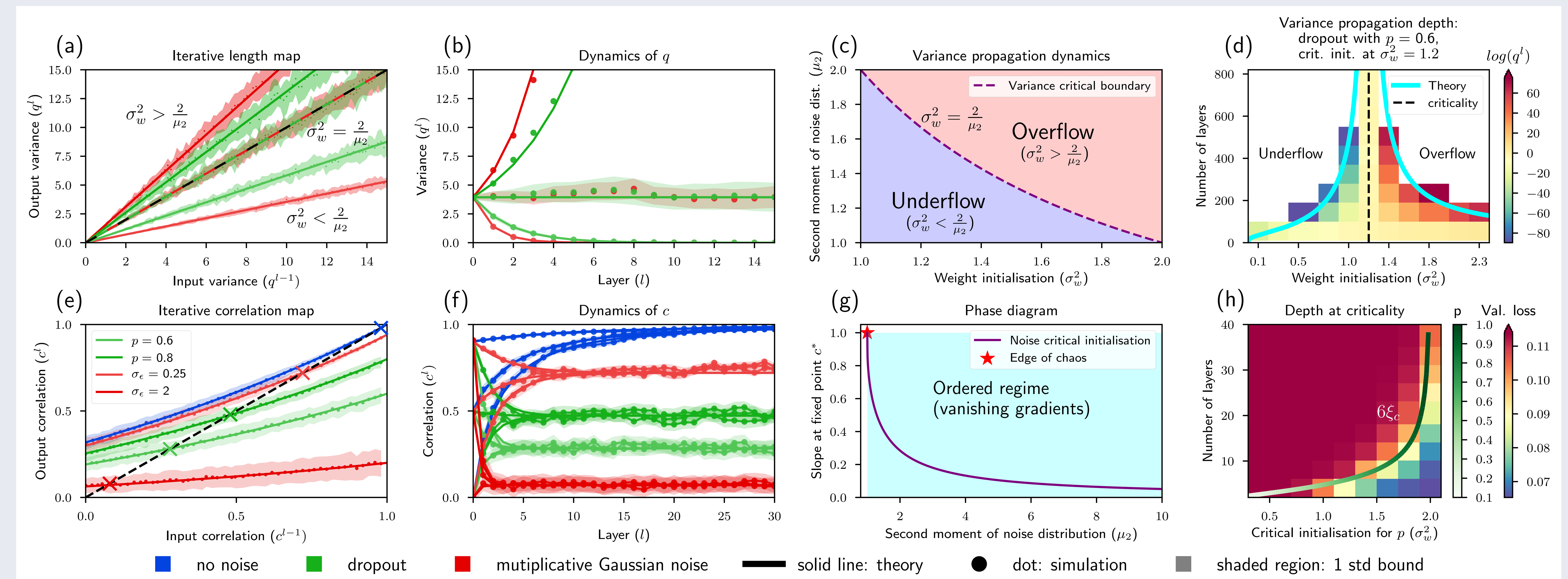


- **Critical initialisation:** $\tilde{q}^* = \sigma_w^2 \left[\tilde{q}^* \mathbb{E}_z(z^2) \odot \mu_2\right] + \sigma_b^2 = \sigma_w^2 \left(\frac{1}{2}\tilde{q}^* \odot \mu_2\right) + \sigma_b^2.$

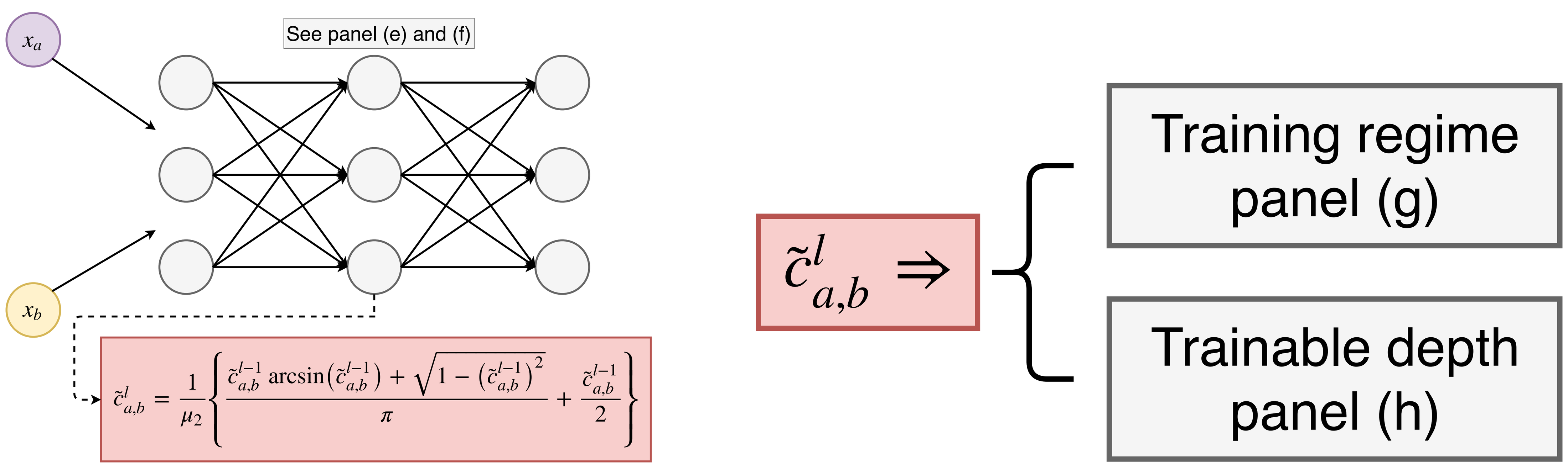
$$\textit{Additive Noise} \implies (\sigma_w, \sigma_b, \mu_2) = (\sqrt{2}, 0, 0)$$
$$\textit{Multiplicative Noise} \implies (\sigma_w, \sigma_b, \mu_2) = \left(\sqrt{2/\mu_2}, 0, \mu_2\right)$$

- **Examples:** Mult Gauss noise: $\mu_2 = \sigma_\epsilon^2 + 1 \implies (\sigma_w, \sigma_b) = \left(\sqrt{2/(\sigma^2+1)}, 0\right)$. Dropout: $\mu_2 = 1/p \implies (\sigma_w, \sigma_b) = (\sqrt{2p}, 0)$ (see panel (c) and (d)).

Experiments



Limits to information flow



- **Training regime:** At the fixed point correlation c^* the slope of the correlation map is

$$\chi(c^*) = \frac{1}{\mu_2}\left(\frac{\arcsin(c^*)}{\pi} + \frac{1}{2}\right)$$

If $\chi(c^*) < 1$, the network is in the *ordered* regime of signal propagation where gradients tend to vanish during the backward pass [2]. Any amount of noise induced regularisation pushes the network into this ordered regime (see panel (g)).

- **Depth scales:** $\xi_c = -1/\log\left[\chi(c^*)\right]$. The value $6\xi_c$ seems to be able to predict feasible depths for trainability [2] and generalisation (see panel (h)).

Takeaways

- When using multiplicative stochastic regularisation techniques (such as dropout) with ReLU networks, critically initialising the weights and biases ensures that information from the input can reliably flow through the network.
- However, even at critically, noise causes the correlation between signals to decay with increasing depth and as a result gradients may vanish during backpropagation. This limits the depth at which noisy ReLU networks are able to perform well.

References

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

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

To be presented at NIPS 2018 Contact: arnupretorius@gmail.com, elanvanbiljon@gmail.com