Activation function design for deep networks: linearity and effective initialisation

- Michael Murray (University of California, Los Angeles)

- Vinayak Abrol (IIIT Delhi)

- **Jared Tanner** (University of Oxford)

Abstract: The activation function deployed in a deep neural network has great influence on the performance of the network at initialisation, which in turn has implications for training. In this paper we study how to avoid two problems at initialisation identified in prior works: rapid convergence of pairwise input correlations, and vanishing and exploding gradients. We prove that both these problems can be avoided by choosing an activation function possessing a sufficiently large linear region around the origin, relative to the bias variance σ_b of the network's random initialisation. We demonstrate empirically that using such activation functions leads to tangible benefits in practice, both in terms test and training accuracy as well as training time. Furthermore, we observe that the shape of the nonlinear activation outside the linear region appears to have a relatively limited impact on training. Finally, our results also allow us to train networks in a new hyperparameter regime, with a much larger bias variance than has previously been possible.

References:

[1] M. Murray, V. Abrol, and J. Tanner Activation function design for deep networks: linearity and effective initialisation. *Applied and Computational Harmonic Analysis*, accepted.

[2] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli Exponential expressivity in deep neural networks through transient chaos. *International conference on neural information processing (NIPS)* , 3368–3376, 2016.

[3] S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein Deep information propagation. *International conference on learning and representations (ICLR)*, 2017.