Deep Neural Network Initialization with Sparsity Inducing Activations

Ilan Price[1,2], Nicholas Daultry Ball [1], Adam C. Jones[1], Samuel C.H. Lam[1], & Jared Tanner[1]
[1]Mathematical Institute, University of Oxford, UK, [2]Alan Turing Institute, UK
{ilan.price, nicholas.daultryball, adam.c.jones, samuel.lam, jared.tanner}@maths.ox.ac.uk

UNIVERSITY OF OXFORD
Mathematical Institute

Outline

Motivation: it is more efficient to compute the "forward pass" of a sparse neural network.

We investigate the training of neural networks with activation functions that induce highly sparse hidden layer outputs throughout both training and inference.

Our approach naturally combine with other standard post-processing procedures to sparsify a network, e.g. weight pruning and quantisation [1] by subselecting only a portion of the weight matrix active for that input.

Edge of chaos (EoC) analysis

A feed-forward neural network could be expressed as the output of the ℓ-th function in the function sequence $\mathbf{h}^\ell(\mathbf{x})$, where

$$h_j^\ell(\mathbf{x}) = \sum_{i=1}^{N_\ell} W_{ij}^\ell \phi(h_i^{\ell-1}(\mathbf{x})) + b_j^\ell. \qquad (1)$$

Assume the weights W_{ij}^ℓ are i.i.d. $\mathsf{N}(0, \sigma_w^2)$, and the biases b_j^ℓ are i.i.d. $\mathsf{N}(0, \sigma_b^2)$. As $N_\ell \to +\infty$, the random functions $\mathbf{h}^\ell$ converge to a Gaussian process with variance $q^\ell := \mathsf{Var}(h_i^\ell)$ that satisfies the recursion $q^{\ell+1} = V_\phi(q^\ell)$ [2], where:

$$V_\phi(q^\ell) := \sigma_b^2 + \sigma_w^2 \int (\phi(\sqrt{q^\ell} z))^2 \, \gamma(dz) \qquad (2)$$

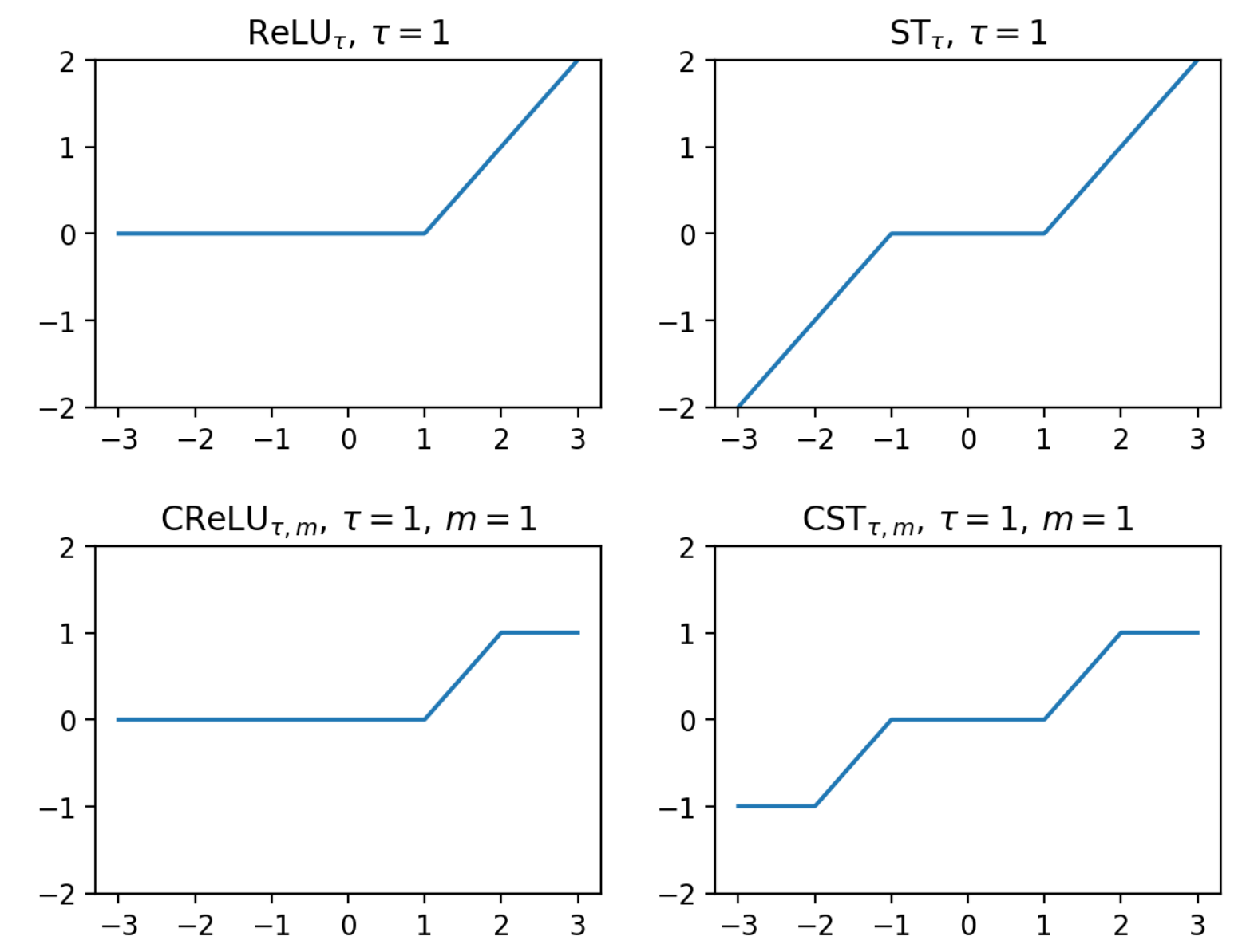
Here γ is the standard Gaussian measure. Assume $q^\ell \equiv q^*$ remain constant. Then the correlation of the Gaussian processes $c_{12}^\ell := \mathsf{corr}(h_i^\ell(\mathbf{x}), h_i^\ell(\mathbf{x}'))$ could also be defined recursively. The sequence of correlations does not converge to the trivial limits (0 or 1) when σ_w^2 is chosen at the "edge of chaos" (EoC), such that:

$$\chi_1 = \sigma_w^2 \int \left[\phi'(\sqrt{q^*} z)\right]^2 \gamma(dz) = 1. \qquad (3)$$

Such initialisation mitigates the problem of vanishing/exploding gradients as well. [3, 4]

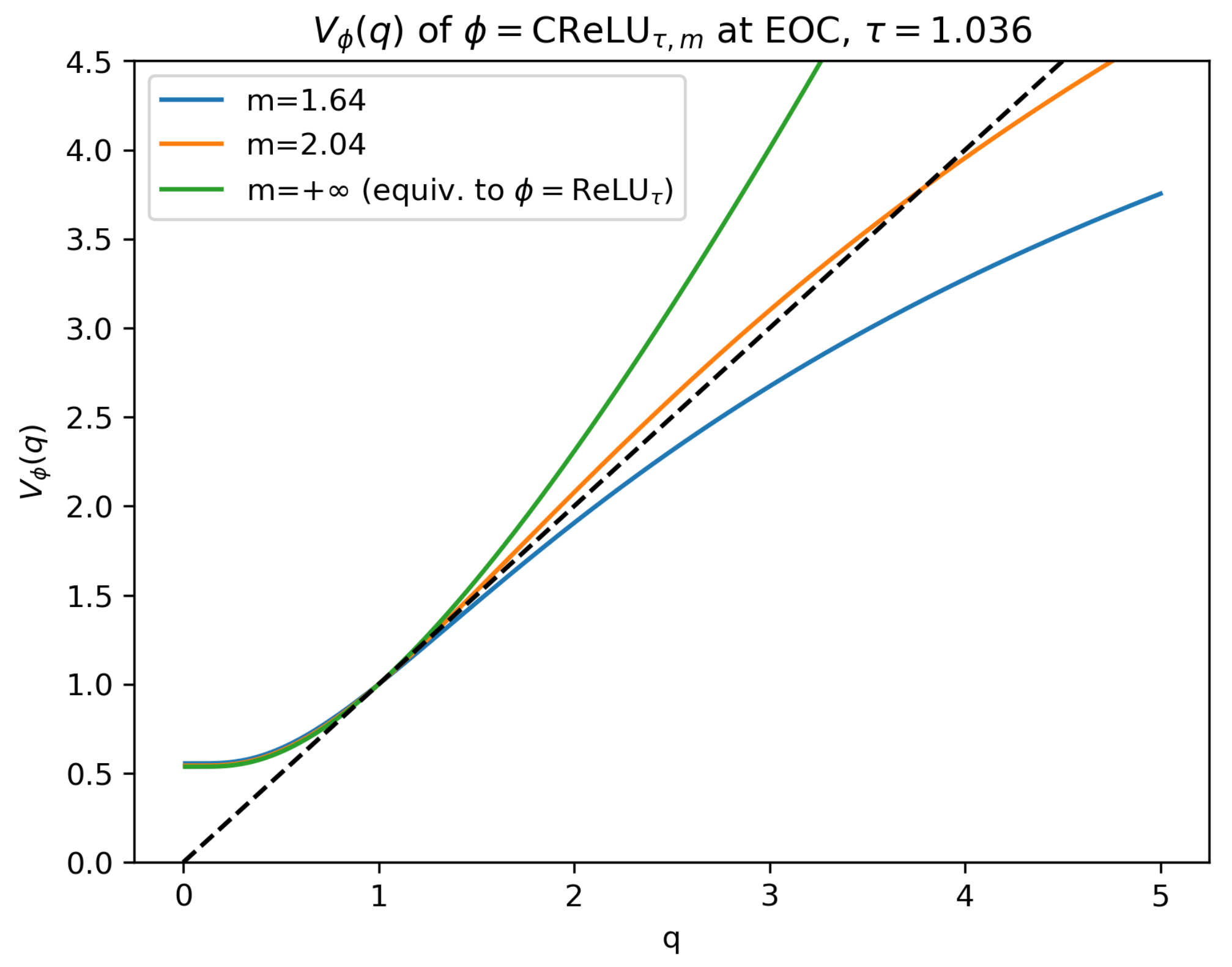
Sparsifying Activation Functions

We choose the functions that are commonly used for signal denoising and compress sensing [5, 6].



Why clipping?

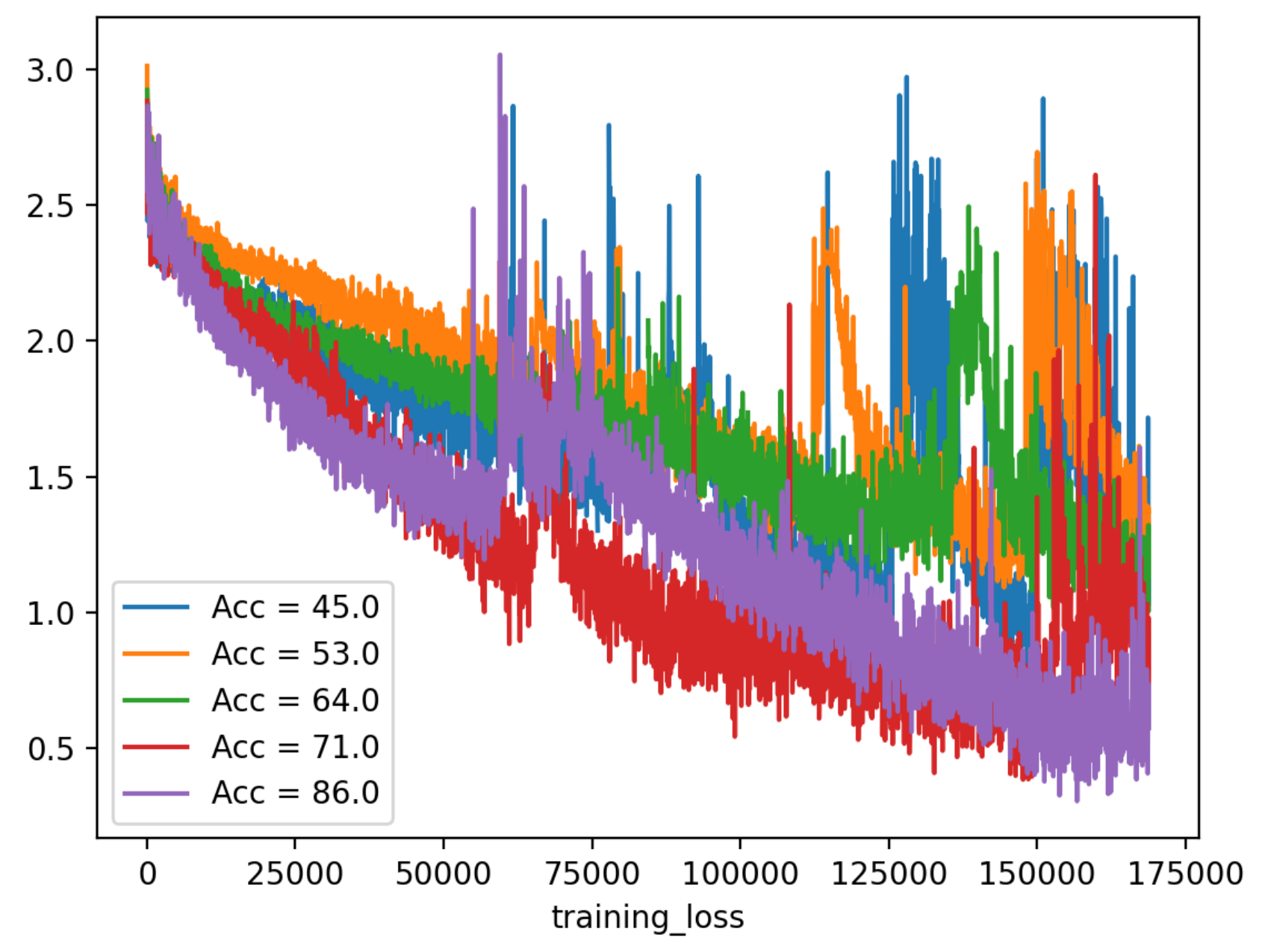
If a feed-forward neural network with $\phi = \mathrm{ReLU}_\tau$ or ST_τ activation functions are initialised at the EoC, then the map $V_\phi(q)$ has a half-stable fixed point q^*, making the assumption of $q^\ell \equiv q^*$ being impractical to achieve. Using a clipped version of the above activation functions, i.e. $\mathrm{CReLU}_{\tau,m}$ or $\mathrm{CST}_{\tau,m}$ could stablise the fixed point.



Failure Modes

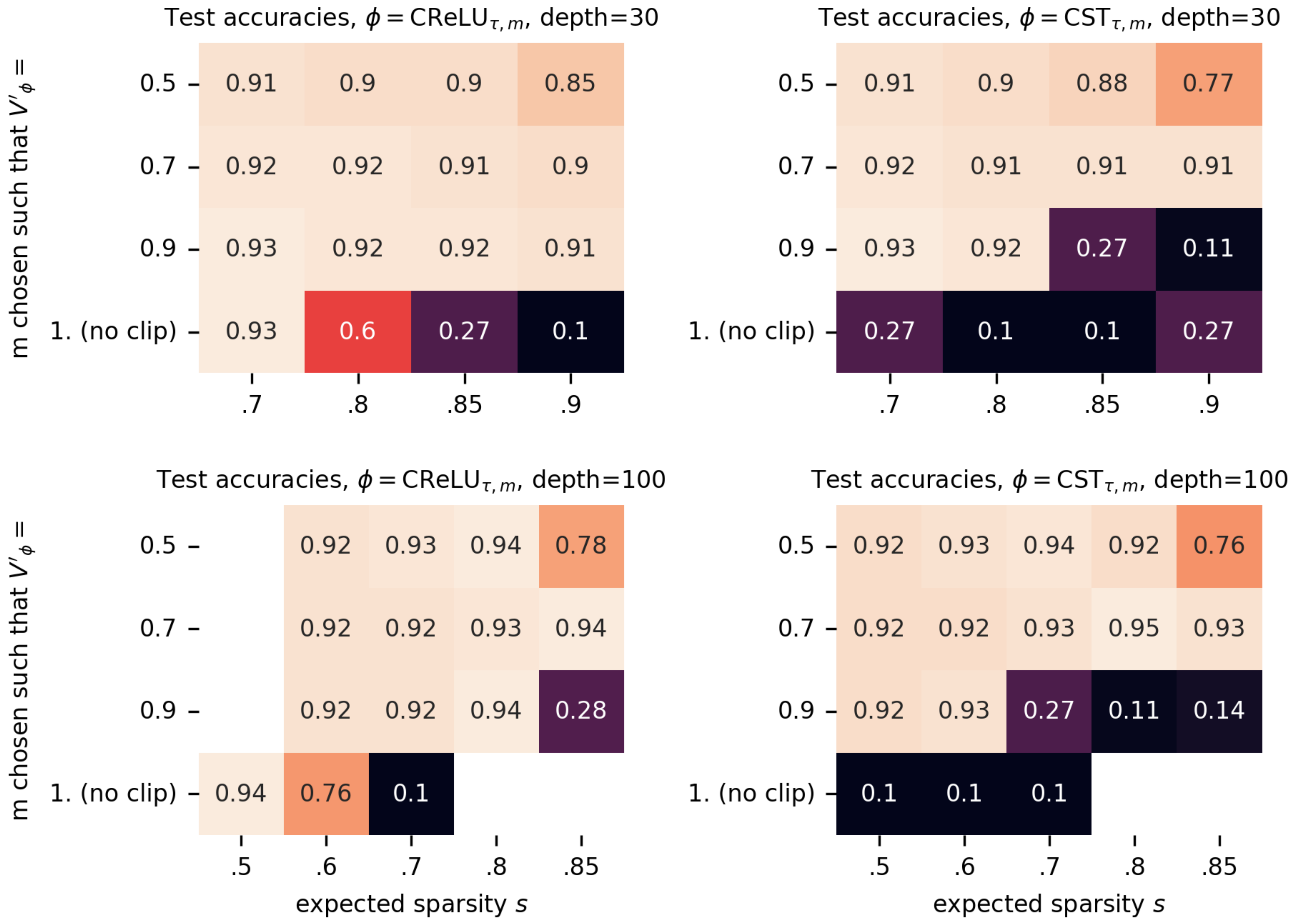
We identify two modes of failures to train a sparse neural network:

- The map $V_\phi(q)$ exhibits more than one fixed points when m is large, so it becomes difficult to initialise at the EoC. **This is captured by the usual vanishing/exploding of gradients.**
- The network is too sparse, so training becomes **unstable** even for small m.



Accuracies could still be retained on sparse networks

Sparse feed-forward neural networks with widths 300 and depths 30 or 100 are trained to classify the digits in the MNIST dataset. The experiments are repeated five times. We report here the mean of the test accuracies.



Test accuracies, $\phi = \mathrm{CReLU}_{\tau,m}$, depth=30

	.7	.8	.85	.9
0.5	0.91	0.9	0.9	0.85
0.7	0.92	0.92	0.91	0.9
0.9	0.93	0.92	0.92	0.91
1. (no clip)	0.93	0.6	0.27	0.1

Test accuracies, $\phi = \mathrm{CST}_{\tau,m}$, depth=30

	.7	.8	.85	.9
0.5	0.91	0.9	0.88	0.77
0.7	0.92	0.91	0.91	0.91
0.9	0.93	0.92	0.27	0.11
1. (no clip)	0.27	0.1	0.1	0.27

Test accuracies, $\phi = \mathrm{CReLU}_{\tau,m}$, depth=100

	.5	.6	.7	.8	.85
0.5		0.92	0.93	0.94	0.78
0.7		0.92	0.92	0.93	0.94
0.9		0.92	0.92	0.94	0.28
1. (no clip)	0.94	0.76	0.1		

Test accuracies, $\phi = \mathrm{CST}_{\tau,m}$, depth=100

	.5	.6	.7	.8	.85
0.5	0.92	0.93	0.94	0.92	0.76
0.7	0.92	0.92	0.93	0.95	0.93
0.9	0.92	0.93	0.27	0.11	0.14
1. (no clip)	0.1	0.1	0.1		

References

[1] Blalock D, Gonzalez Ortiz JJ, Frankle J, Guttag J. What is the State of Neural Network Pruning? In: Dhillon I, Papailiopoulos D, Sze V, editors. Proceedings of Machine Learning and Systems. vol. 2; 2020. p. 129-46. Available from: https://proceedings.mlsys.org/paper_files/paper/2020/file/6c44dc73014d66ba49b28d483a8f8b0d-Paper.pdf.
[2] Poole B, Lahiri S, Raghu M, Sohl-Dickstein J, Ganguli S. Exponential expressivity in deep neural networks through transient chaos. Advances in neural information processing systems. 2016;29.
[3] Schoenholz SS, Gilmer J, Ganguli S, Sohl-Dickstein J. Deep Information Propagation. In: International Conference on Learning Representations; 2017. Available from: https://openreview.net/forum?id=H1W1UN9gg.
[4] Pennington J, Schoenholz S, Ganguli S. The emergence of spectral universality in deep networks. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2018. p. 1924-32.
[5] Donoho DL. De-noising by soft-thresholding. IEEE Transactions on Information Theory. 1995;41(3):613-27.
[6] Foucart S, Rauhut H. A Mathematical Introduction to Compressive Sensing. Applied and Numerical Harmonic Analysis. Birkhäuser; 2013.