RePU is All You Need (Work in Progress)

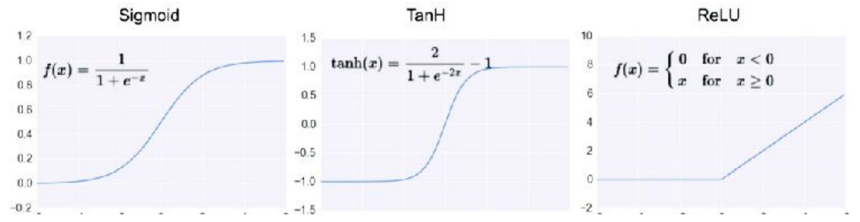
Qiyao Wei, Martin Magill, Luca Herranz-Celotti, Ermal Rrapaj

UNIVERSITY OF TORONTO

VECTOR INSTITUTE

Neural Network Activation Functions

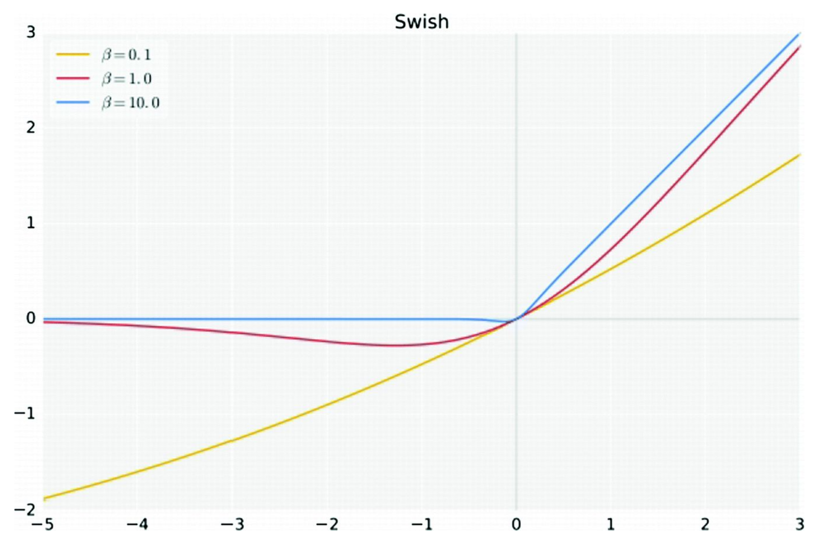
There is an abundance of popular activation functions



But there are also overlooked activations w/ desirable properties

Controlled swish: smooth ReLU with one hyperparameter
RePU: ReLU taken to polynomial power

Differentiable everywhere!



Neural Network Metrics

We assume mean-zero Gaussian initialization w/ variance

$$\mathbb{E}\left[b_i^{(1)} b_j^{(1)}\right] = \delta_{ij} C_b^{(1)}$$

$$\mathbb{E}\left[W_{i_1 j_1}^{(1)} W_{i_2 j_2}^{(1)}\right] = \delta_{i_1 i_2}\delta_{j_1 j_2}\frac{C_W^{(1)}}{n_0}$$

We can calculate some correlators for different neurons

$$\mathrm{E}\left[z_{i;\alpha}^{(1)}\right] = \mathbb{E}\left[b_i^{(1)} + \sum_{j=1}^{n_0} W_{ij}^{(1)} x_{j;\alpha}\right] = 0$$

$$\mathbb{E}\left[z_{i_1;\alpha_1}^{(1)} z_{i_2;\alpha_2}^{(1)}\right] = \mathbb{E}\left[\left(b_{i_1}^{(1)} + \sum_{j_1=1}^{n_0} W_{i_1 j_1}^{(1)} x_{j_1;\alpha_1}\right)\left(b_{i_2}^{(1)} + \sum_{j_2=1}^{n_0} W_{i_2 j_2}^{(1)} x_{j_2;\alpha_2}\right)\right]$$

$$= \delta_{i_1 i_2}\left(C_b^{(1)} + C_W^{(1)}\frac{1}{n_0}\sum_{j=1}^{n_0} x_{j;\alpha_1} x_{j;\alpha_2}\right) = \delta_{i_1 i_2} G_{\alpha_1 \alpha_2}^{(1)}$$

$$\mathbb{E}\left[z_{i_1;\alpha_1}^{(2)} z_{i_2;\alpha_2}^{(2)} z_{i_3;\alpha_3}^{(2)} z_{i_4;\alpha_4}^{(2)}\right]\Big|_{\text{connected}}$$
$$= \frac{1}{n_1}\left[\delta_{i_1 i_2}\delta_{i_3 i_4} V_{(\alpha_1\alpha_2)(\alpha_3\alpha_4)}^{(2)} + \delta_{i_1 i_3}\delta_{i_2 i_4} V_{(\alpha_1\alpha_3)(\alpha_2\alpha_4)}^{(2)} + \delta_{i_1 i_4}\delta_{i_2 i_3} V_{(\alpha_1\alpha_4)(\alpha_2\alpha_3)}^{(2)}\right]$$

Categorizing Activation Functions

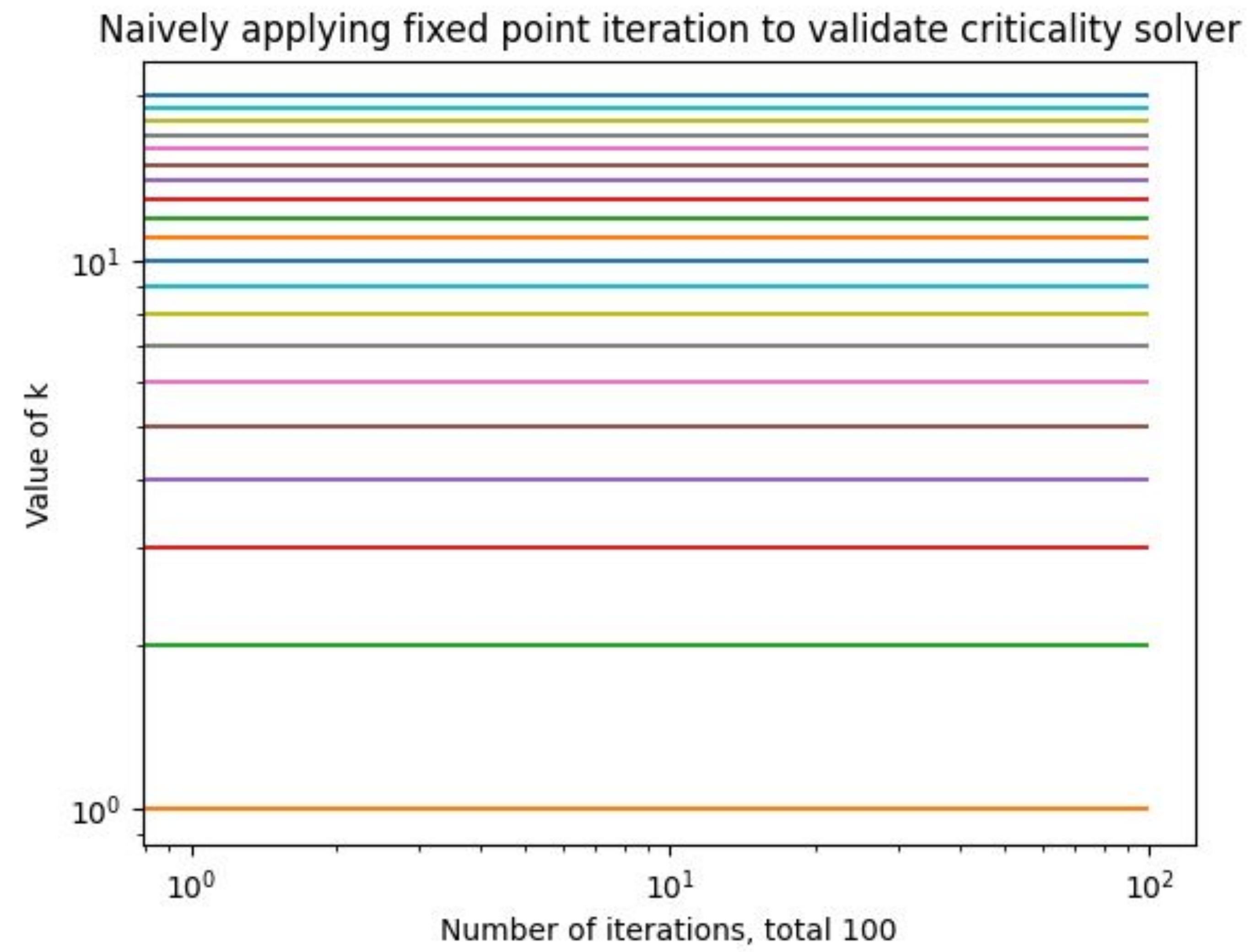
Most activation functions can be grouped into just a few classes:

Scale-Invariant Activations: ReLU
No Criticality: sigmoid, softplus, nonlinear monomials
K = 0 Universality Class: tanh, sin
Half-Stable Universality Classes: SWISH and GELU

This is reminiscent of the vanishing/exploding gradient problem in deep neural networks. We want our activation functions to put the neural net on the “critical sweet spot”.

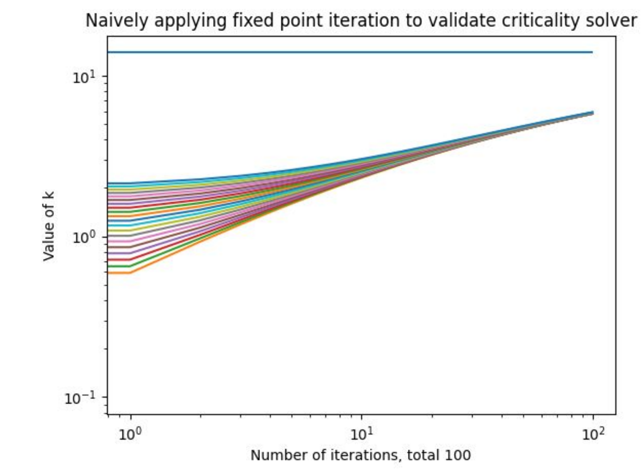
What Does This Say About Activations?

ReLU has “a line of fixed points”

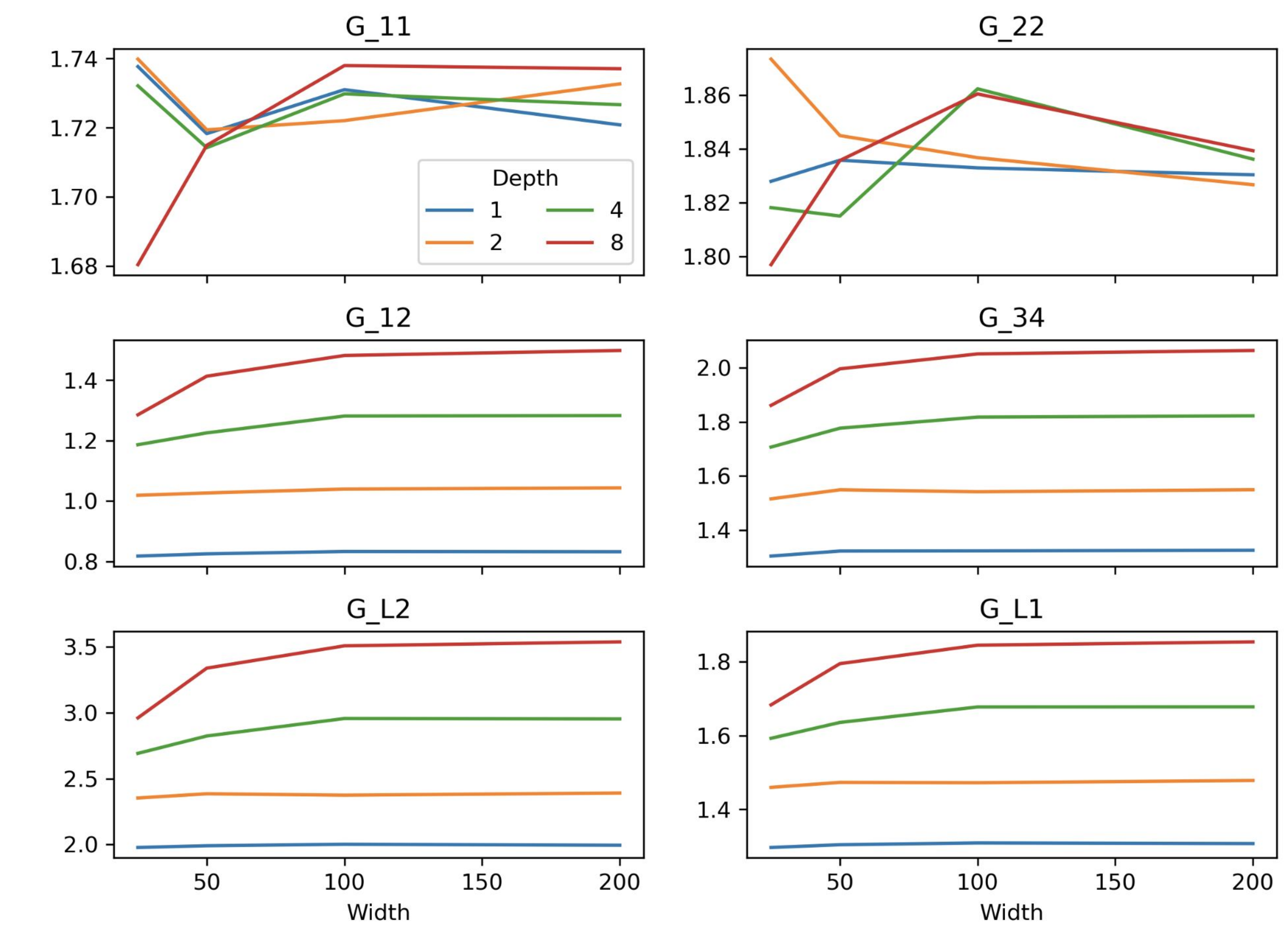


Experiments

1) Controlled swish at Criticality



2) ReLU Activation in Action



References

Roberts, D. A., Yaida, S., & Hanin, B. (2021). The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*.