Discovery. Diversity. Distinction.

TAct: Optimal search through activation function space

Mario Banuelos, Heyley Gatewood, Samuel Hood, Jonathan Scott, and David Uminsky

mbgmath.com | @mbanuelos22

MBG

Acknowledgements

- **Heyley Gatewood**, Stetson University
- **Samuel Hood**, Morehouse College
- **Jonathan Scott**, Stetson University
- Mercedes Franco, Queensborough Community College
- David Uminsky, University of San Francisco (USF)
- Yannet Interian, USF

mbanuelos22@csufresno.edu

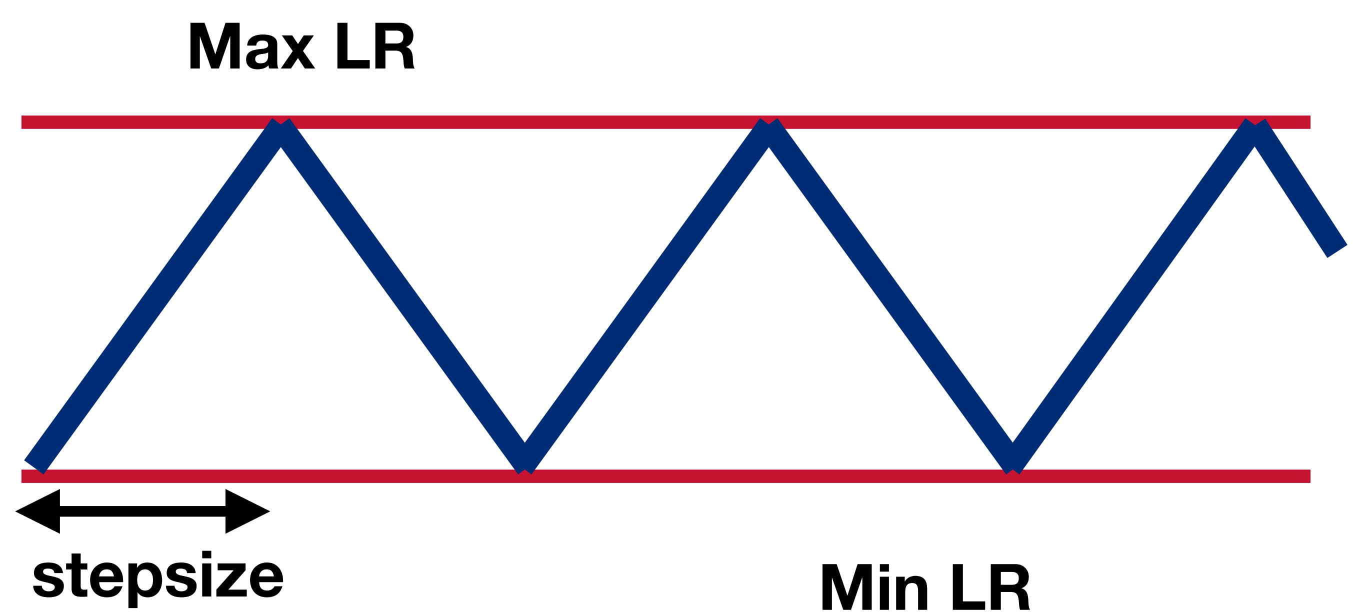
The Problem

- Deep learning methods addressing approximation of data should be generalizable.
- Hyperparameters and activation functions help accomplish this task.

mbanuelos22@csufresno.edu

The Problem

- Deep learning methods addressing approximation of data should be generalizable.
- Hyperparameters and activation functions help accomplish this task.
- Recent work has addressed this issue with learning rates [1].



[1] Smith, L. N. *Cyclical learning rates for training neural networks*. In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, pp. 464-472. IEEE, 2017.

Activation Functions

- Many activation functions have been proposed (too many to enumerate!)

- *Traditional approach:*
  1) Fix a model
  2) Exhaustively incorporate different activation functions
  3) Report the highest accuracy model

Activation Functions

- Many activation functions have been proposed (too many to enumerate!)
- *Traditional approach:*
  1) Fix a model
  2) Exhaustively incorporate different activation functions
  3) Report the highest accuracy model
- **Why not let the problem inform the choice of activation function?**

Minimizing the L_2 distance

- A norm in C_2 in the interval *[a,b]* is defined by

$$\| f \| = \left(\int_a^b f^2(t)\, d \right)^{1/2}$$

mbanuelos22@csufresno.edu

Minimizing the L_2 distance

- A norm in C_2 in the interval *[a,b]* is defined by

$$\| f \| = \left(\int_a^b f^2(t)\, d \right)^{1/2}$$

- A distance between functions *f* and *g* becomes

$$d(f, g) = \left(\int_a^b [g(t) - f(t)]^2\, d \right)^{1/2}$$

mbanuelos22@csufresno.edu

Minimizing the L_2 distance

- A norm in C_2 in the interval *[a,b]* is defined by

$$\| f \| = \left(\int_a^b f^2(t)\, d \right)^{1/2}$$

- A distance between functions *f* and *g* becomes

$$d(f,g) = \left(\int_a^b [g(t) - f(t)]^2\, d \right)^{1/2}$$

- **Our Approach: Using a generalized function, minimize distance to *existing* activation functions.**

Formulating TAct

- We propose a two-parameter, trainable Tanh activation function, which we call **TAct**.

- Exactly contains classic functions such as Tanh, Sigmoid and more recently Swish.

- Approximates functions like ReLu arbitrarily closely.

Formulating TAct

- To create this parameter space, we form a convex hull of nonlinear interpolations between these three activation functions:

$$\mathrm{TAct}(x) := \left(\frac{\mu+1}{6}x + \frac{2-\mu}{6}\right)\left(\tanh\left(\frac{\gamma+4}{6}x\right) + 1\right).$$

- We initialize parameters from a uniform distribution in [-1,1] x [-1,1].

mbanuelos22@csufresno.edu

Formulating TAct

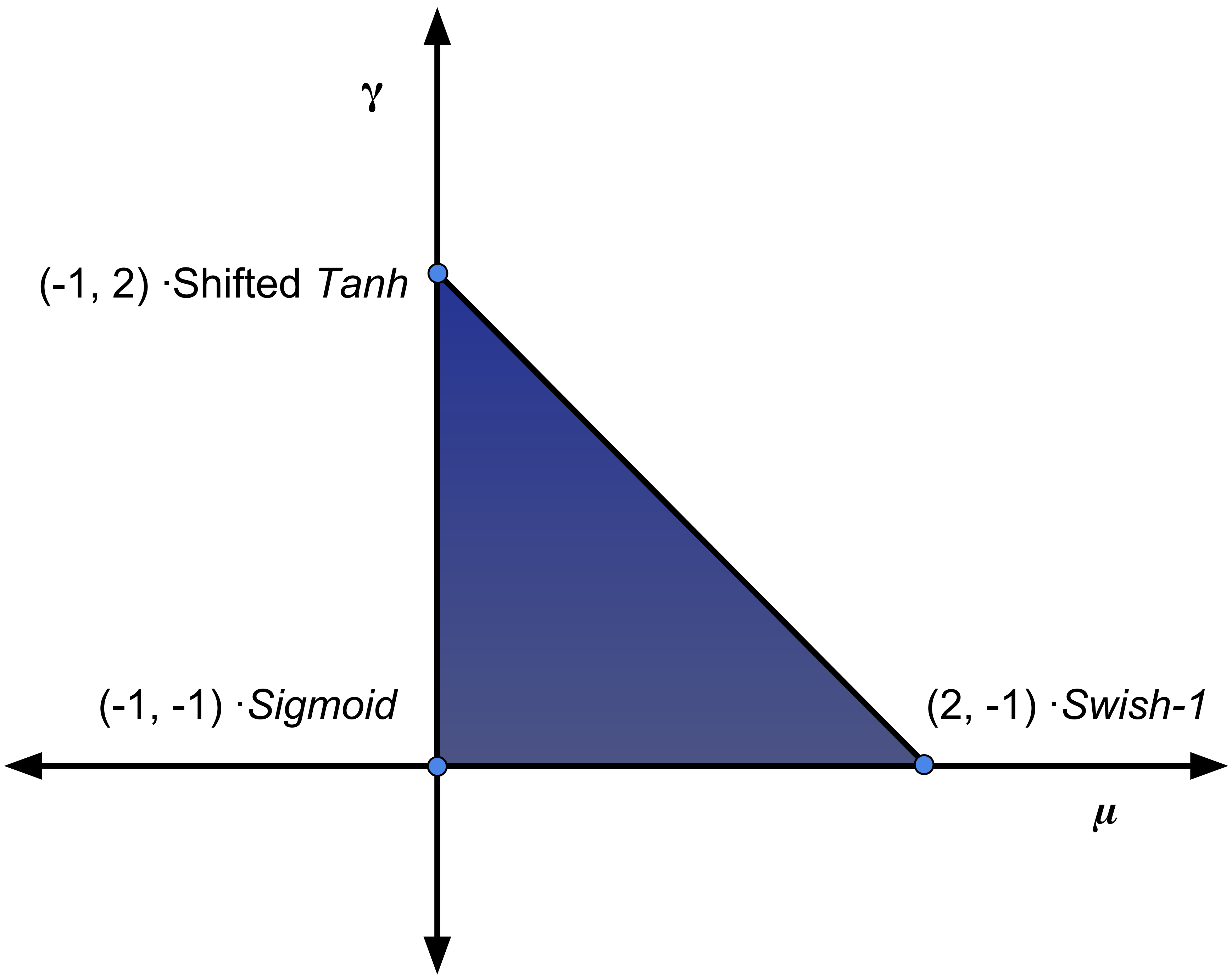
- To create this parameter space, we form a convex hull of nonlinear interpolations between these three activation functions:

$$\mathrm{TAct}(x) := \left(\frac{\mu+1}{6}x + \frac{2-\mu}{6}\right)\left(\tanh\left(\frac{\gamma+4}{6}x\right)+1\right).$$

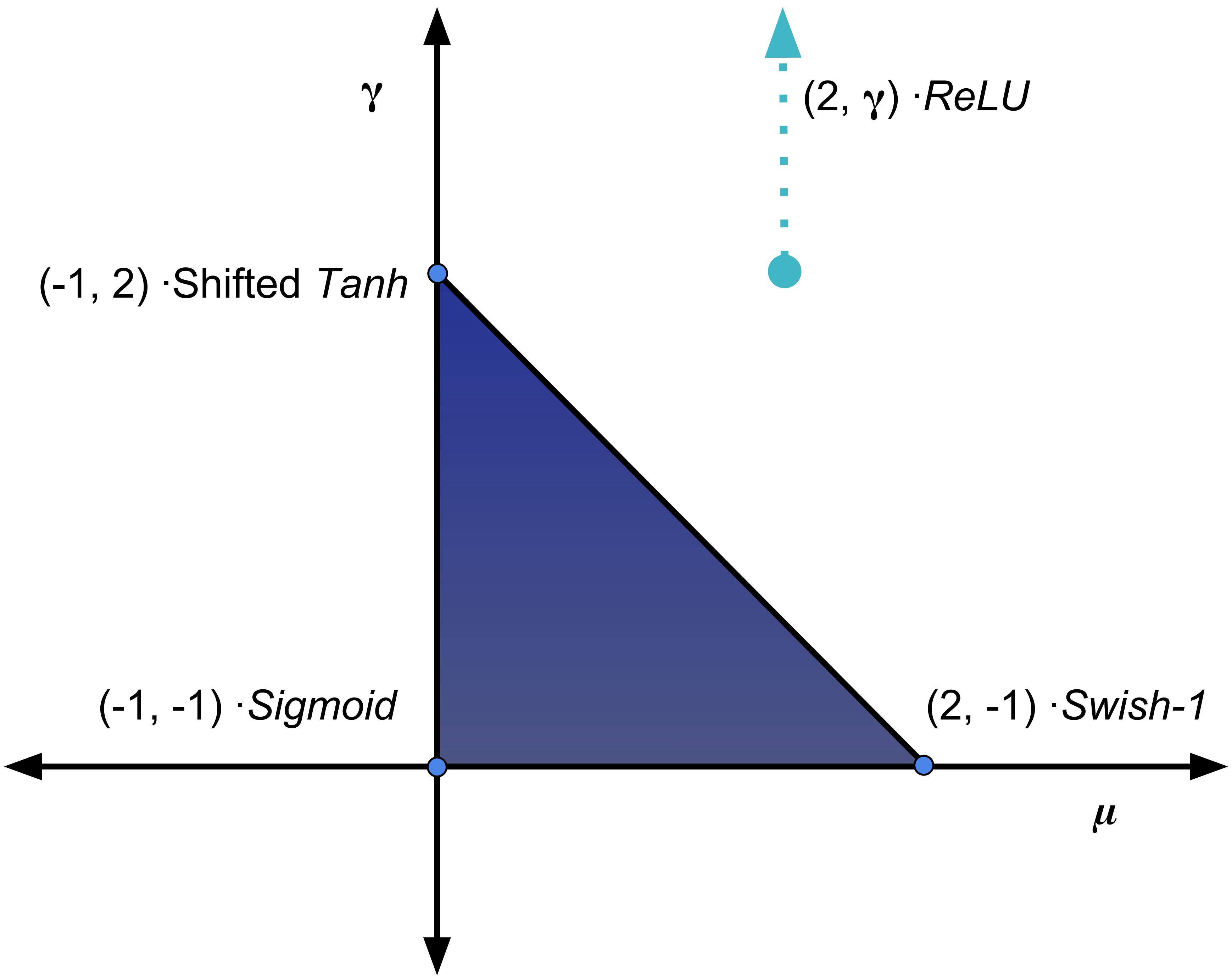
- We initialize parameters from a uniform distribution in [-1,1] x [-1,1].

mbanuelos22@csufresno.edu

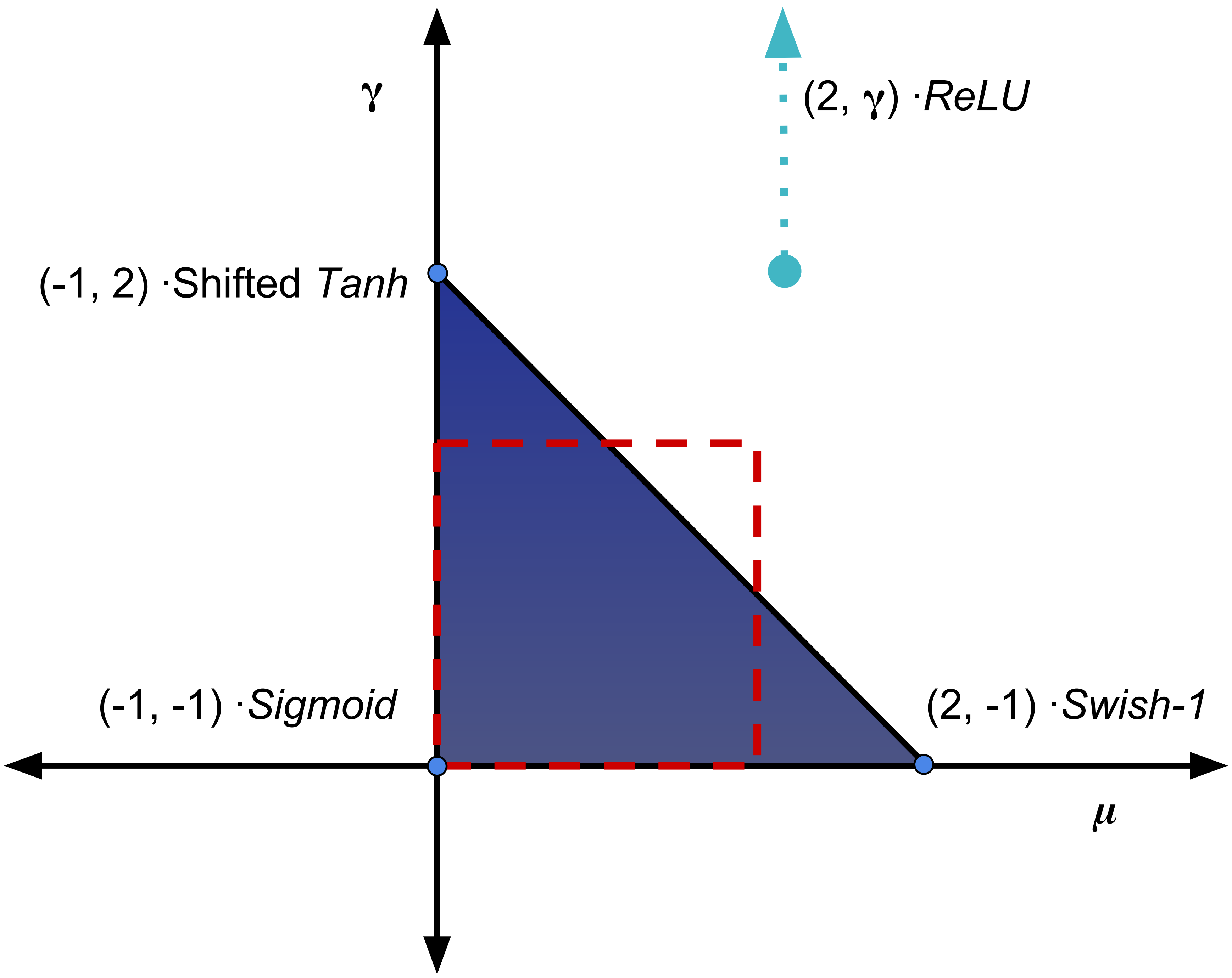
Visualizing TAct



Visualizing TAct



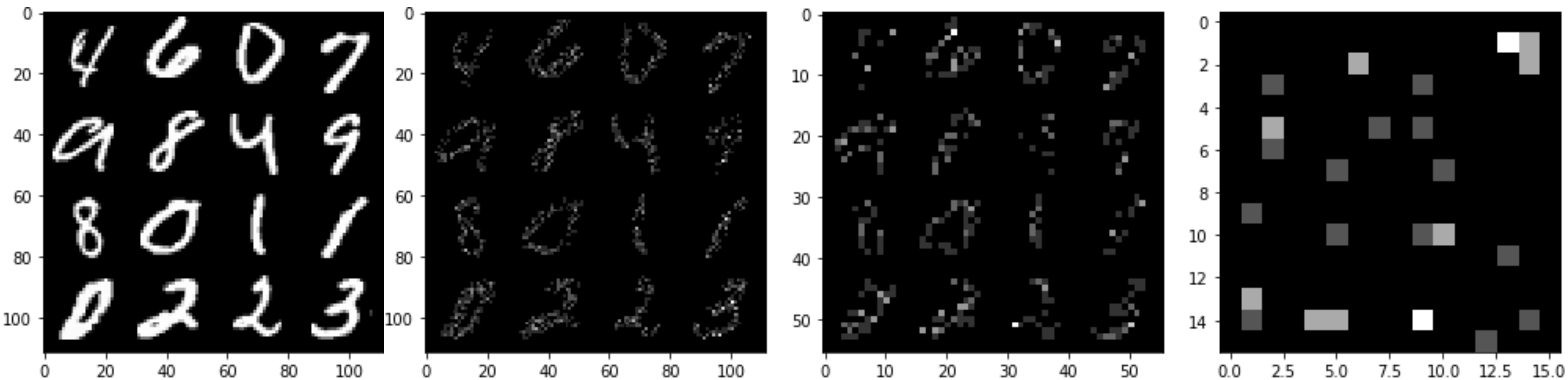
Visualizing TAct



Experiments

- Initial exploration with MNIST with poisson noise with fixed learning rates.
- Lower resolution is upsampled to 28 x 28 and then classified with LeNet-5 architecture.

Poisson MNIST



Poisson MNIST

Activation	28 x 28	14 x 14	7 x 7	4 x 4
ReLU	0.0216	0.0697	0.1887	0.3886
TAct	**0.0190**	**0.0519**	**0.1719**	**0.3717**

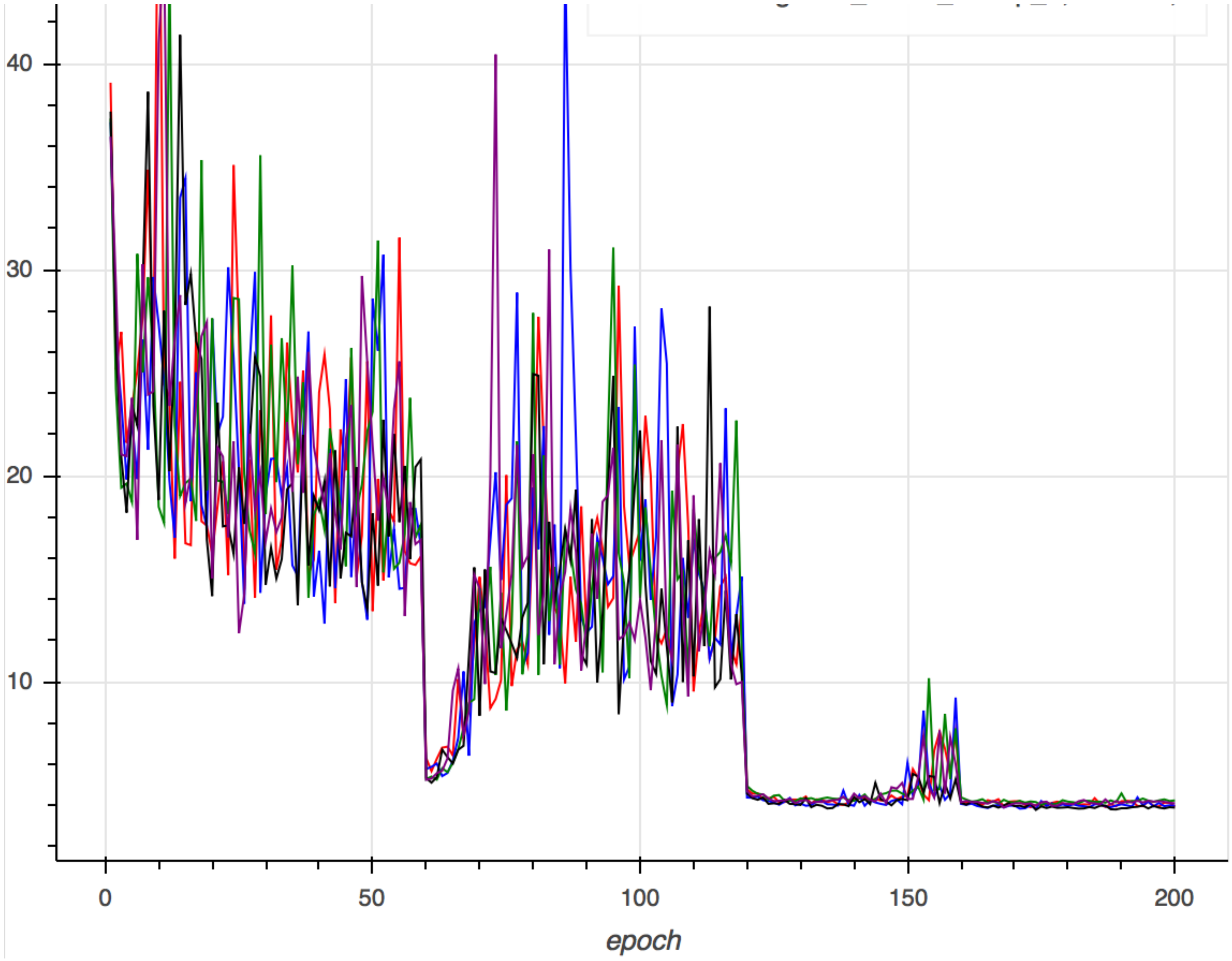
mbanuelos22@csufresno.edu

CIFAR Experiments

- Incorporate wide residual networks (WRN 28-10), for 200 epochs, using TAct.

WRN 28-10	Test Acc.
LReLU	95.6
Softplus	94.9
ReLU	95.3
Swish-1	95.3
TAct	**95.94**

CIFAR-10 Test Error

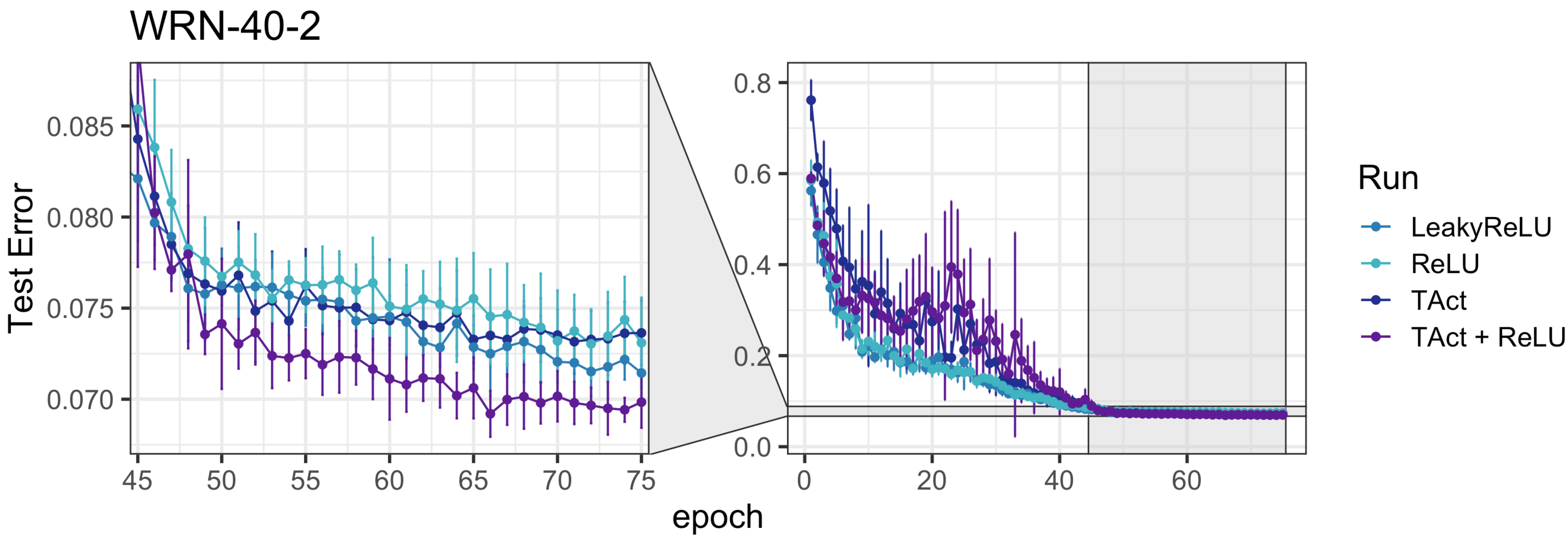


CIFAR Experiments

- Updated parameters: triangular learning rates, data augmentation in CIFAR-10 and CIFAR-100

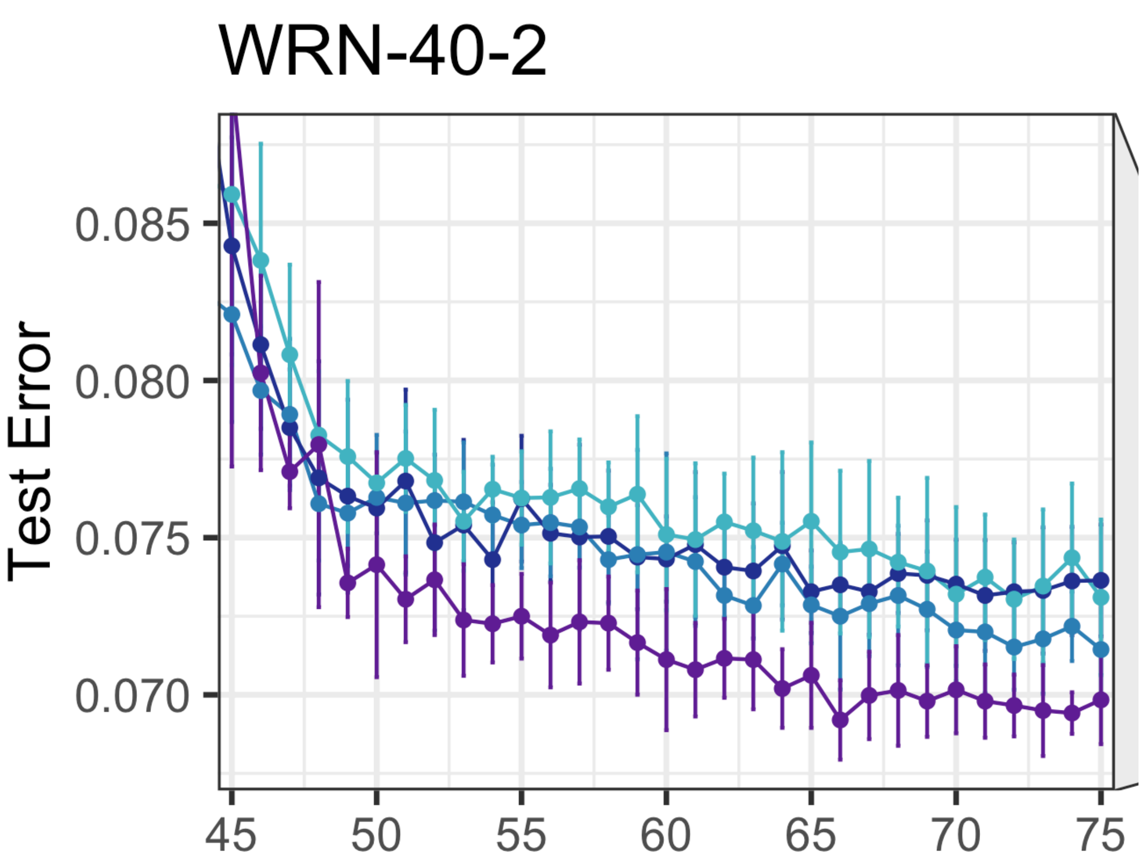
Wide Residual Networks	**DarkNet Architectures**
$N = [\,3, 4, 5, 6\,], k = 2$	Darknet-39
$N = 4, k = 6$ — (22-6)	Darknet-53
$N = 6, k = 2$ — (40-2)	

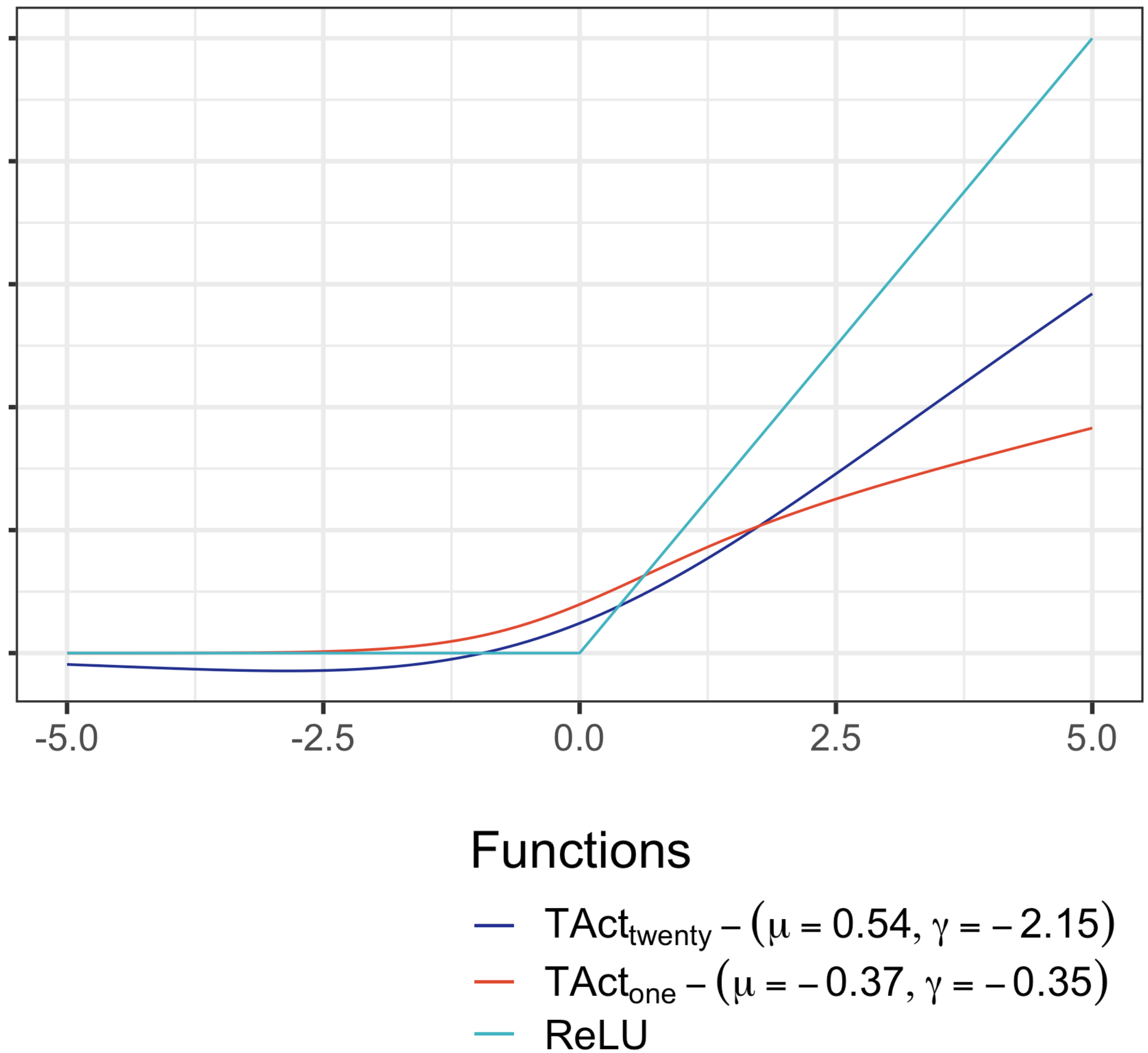
CIFAR-10



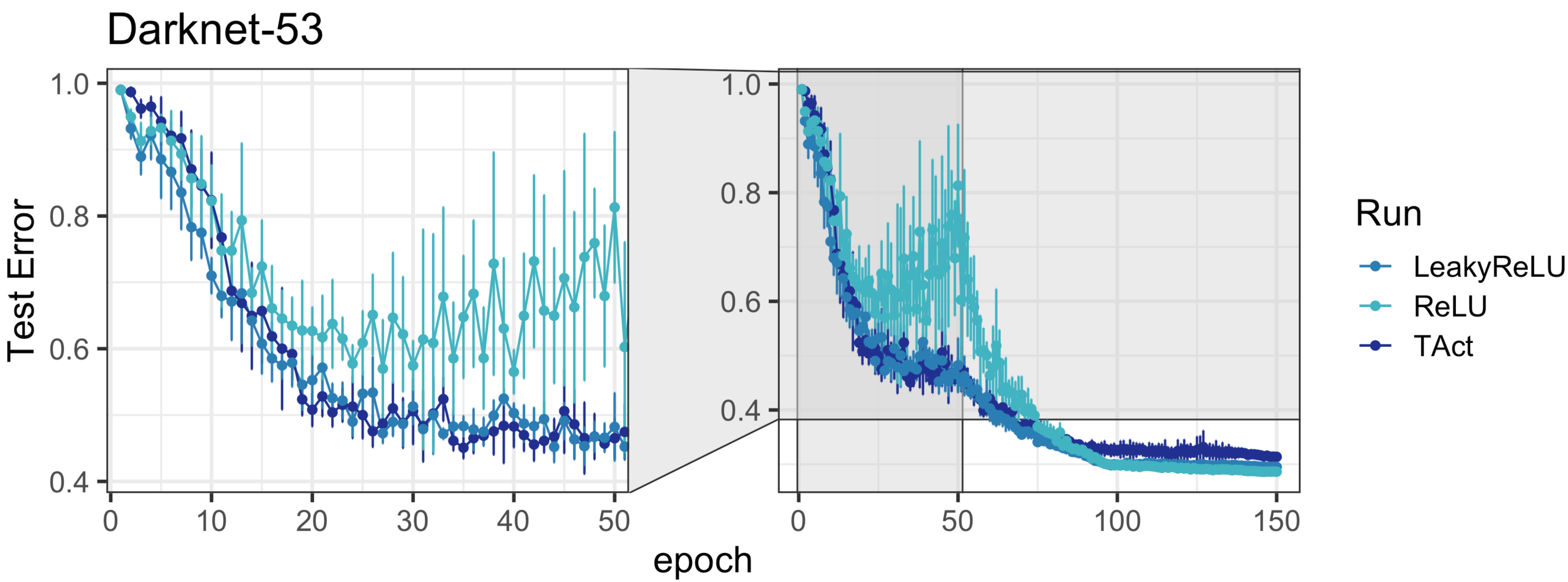
mbanuelos22@csufresno.edu

CIFAR-10


WRN-40-2
Test Error
0.085
0.080
0.075
0.070
45
50
55
60
65
70
75

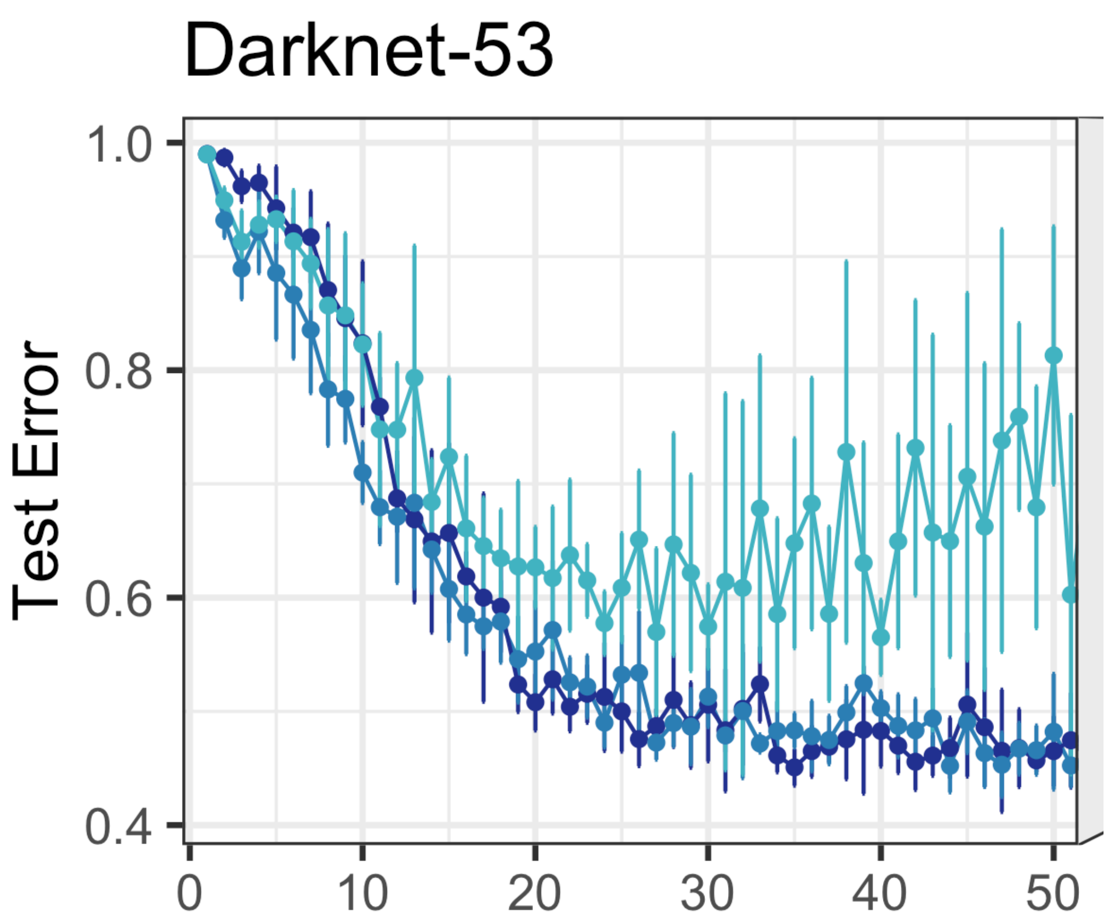

-5.0
-2.5
0.0
2.5
5.0
Functions
$\mathrm{TAct}_{twenty} - (\mu = 0.54, \gamma = -2.15)$
$\mathrm{TAct}_{one} - (\mu = -0.37, \gamma = -0.35)$
ReLU

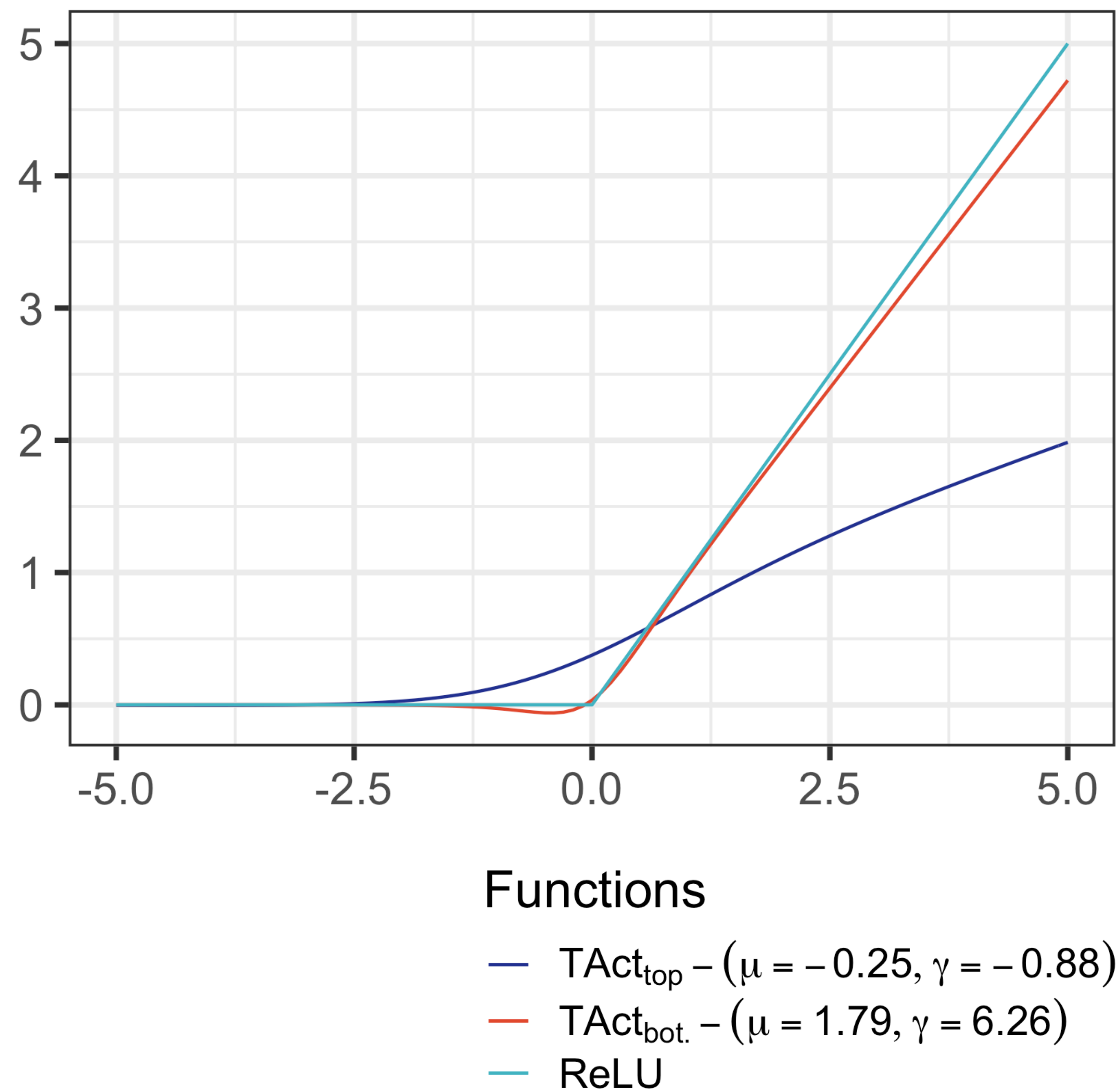
CIFAR-100



Darknet-53
Test Error
1.0
0.8
0.6
0.4
0
10
20
30
40
50
0
50
100
150
epoch
Run
LeakyReLU
ReLU
TAct

CIFAR-100





mbanuelos22@csufresno.edu

Conclusions

- By letting the data drive the choice of activation functions, we achieve competitive test error rates when compared to other popular activation functions.

- We are currently conducting a more thorough comparison across more activation functions and architectures.

MSRI
Mathematical Sciences
Research Institute

NATIONAL SECURITY AGENCY
UNITED STATES OF AMERICA

LXAI

NSF

SLOAN FOUNDATION
ALFRED P. SLOAN FOUNDATION
1934
Alfred P. Sloan
FOUNDATION

FRESNO
STATE
Discovery. Diversity. Distinction.

References

[1] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Tech. rep. Citeseer, 2009.

[2] K. Jarrett, K. Kavukcuoglu, Y. LeCun, et al. "What is the best multi-stage architecture for object recognition?" In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE. 2009, pp. 2146–2153.

[3] V. Nair and G. E. Hinton. "Rectified linear units improve restricted boltzmann machines". In: Proceedings of the 27th international conference on machine learning (ICML-10). 2010, pp. 807–814.

[4] A. L. Maas, A. Y. Hannun, and A. Y. Ng. "Rectifier nonlinearities improve neural network acoustic models". In: Proc. icml. Vol. 30. 1. 2013, p. 3.

[5] K. He et al. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: Proceedings of the IEEE international conference on computer vision. 2015, pp. 1026–1034.

[6] D.A. Clevert, T. Unterthiner, and S. Hochreiter. "Fast and accurate deep network learning by exponential linear units (elus)". In: arXiv preprint arXiv:1511.07289 (2015).

[7] G. Klambauer et al. "Self-normalizing neural networks". In: Advances in Neural Information Processing Systems. 2017, pp. 971–980.

[8] P. Ramachandran, B. Zoph, and Q. V. Le. "Searching for Activation Functions". In: CoRR abs/1710.05941 (2017). arXiv: 1710.05941. url: http://arxiv.org/abs/1710.05941.

[9] H. Chung, S. J. Lee, and J. G. Park. "Deep neural network using trainable activation functions". In: Neural Networks (IJCNN), 2016 International Joint Conference on. IEEE. 2016, pp. 348–352.

[10] J. Redmon and A. Farhadi. "Yolov3: An incremental improvement". In: arXiv preprint arXiv:1804.02767 (2018).