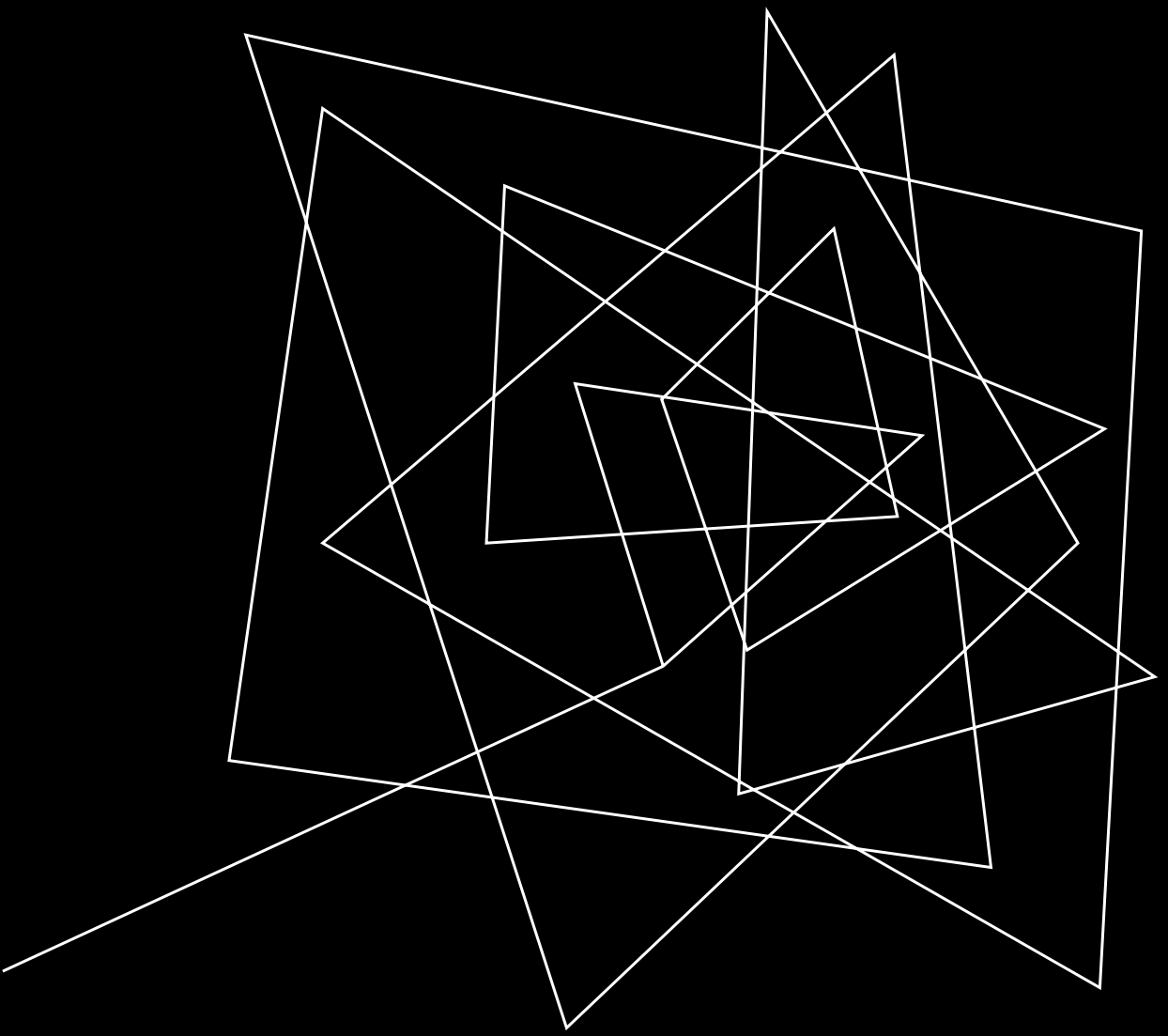
LOSS LANDSCAPES ARE ALL YOU NEED: NEURAL NETWORK GENERALIZATION CAN BE EXPLAINED WITHOUT THE IMPLICIT BIAS OF GRADIENT DESCENT

Ping-yeh Chiang, Renkun Ni, David Y. Miller, Arpit Bansal, Jonas Geiping, Micah Goldblum, Tom Goldstein

UNIVERSITY OF MARYLAND

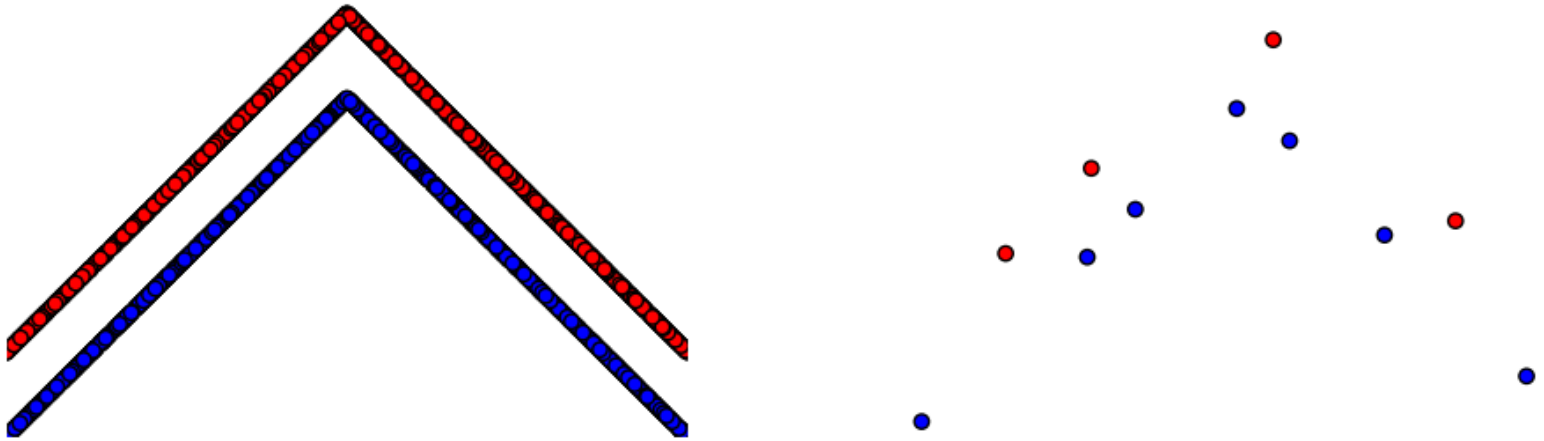
MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

NYU

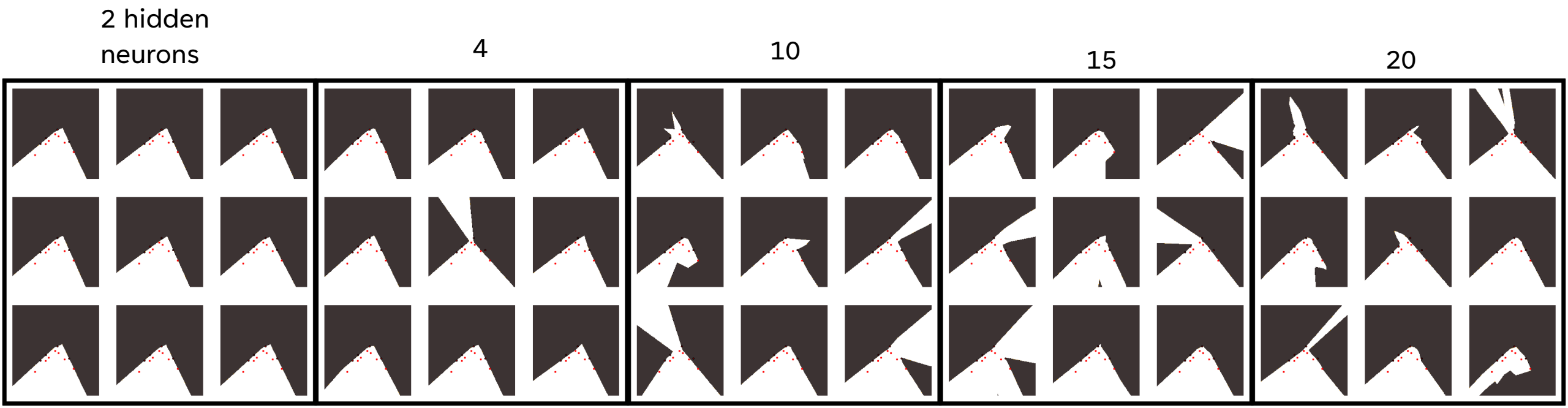


WHY DO NEURAL NETWORKS GENERALIZE?

AN ILLUSTRATIVE TOY EXAMPLE

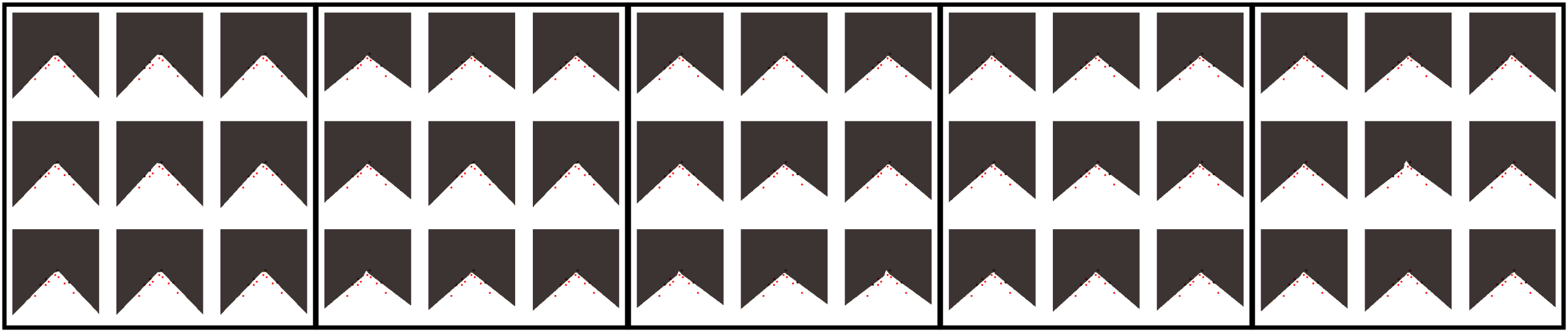


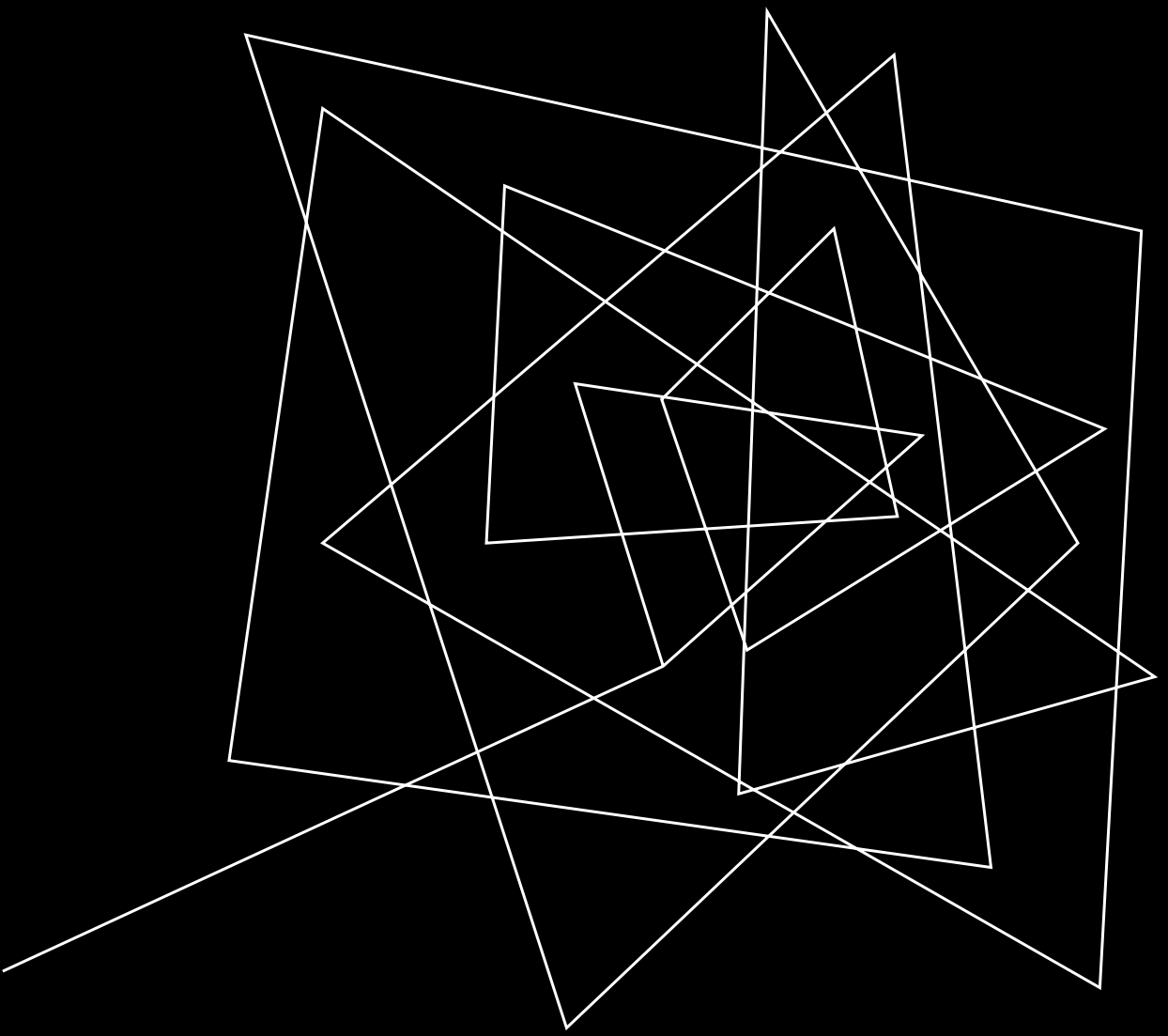
IN OVERPARAMETRIZED MODELS, BAD MINIMA EXIST...



11 training examples are fit with SGD with a “poisoned” loss to fail on five test examples

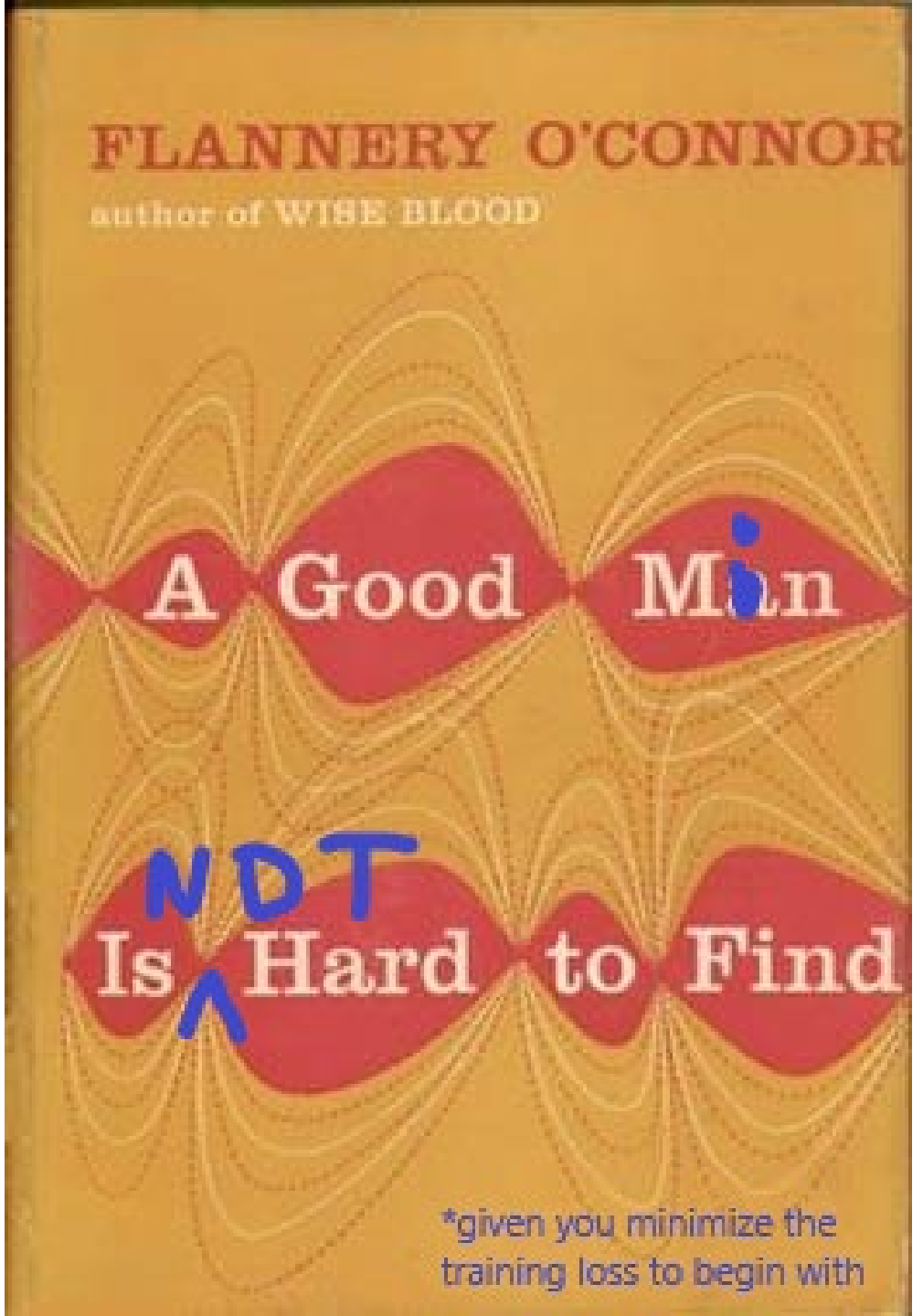
...BUT YOU WOULDN'T KNOW IT FROM USING SGD-TRAINED NETWORKS





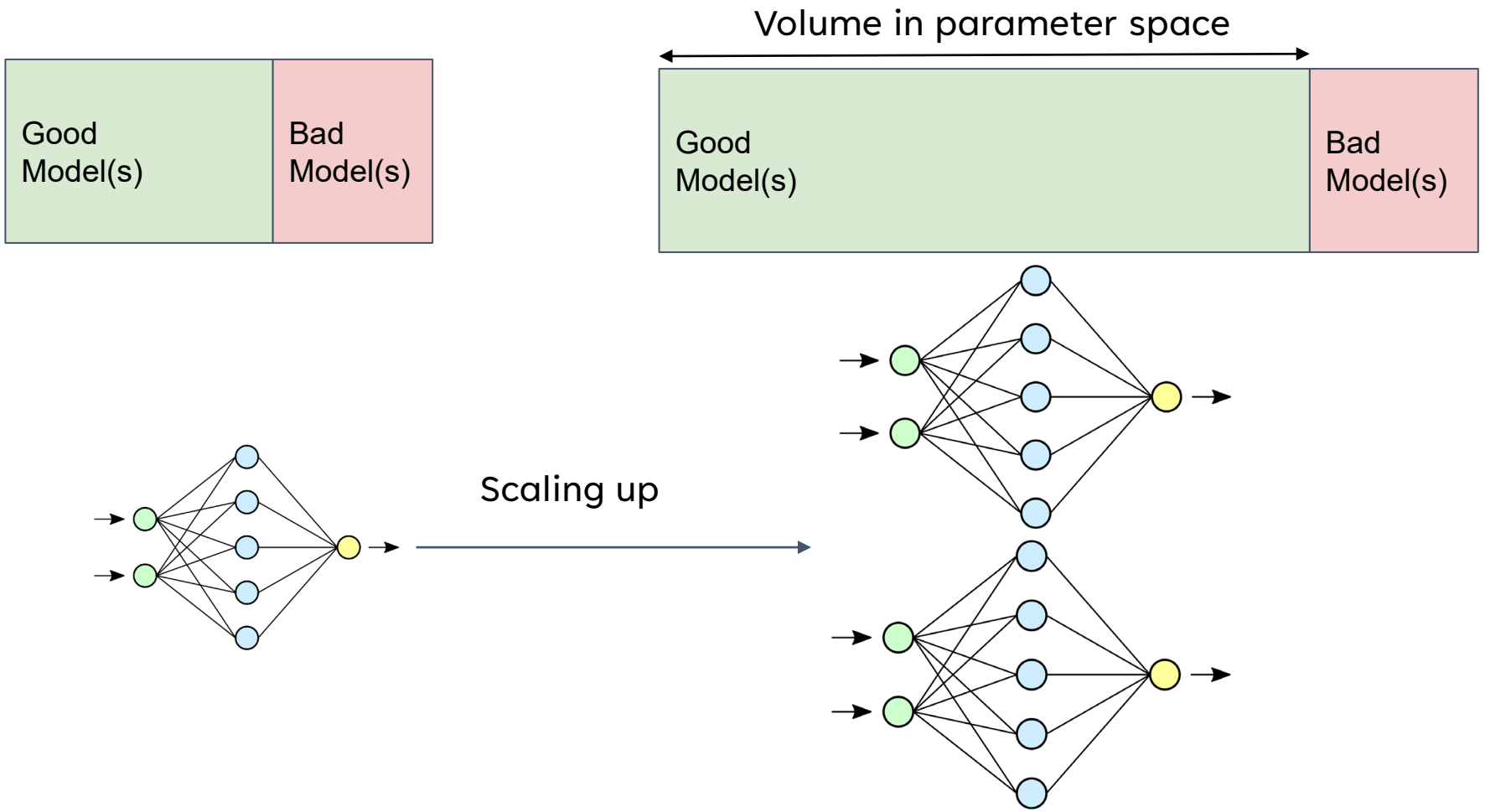
IS SGD SPECIAL?

AN ALTERNATIVE HYPOTHESIS



https://en.wikipedia.org/wiki/File:AGoodManIsHardToFind.jpg

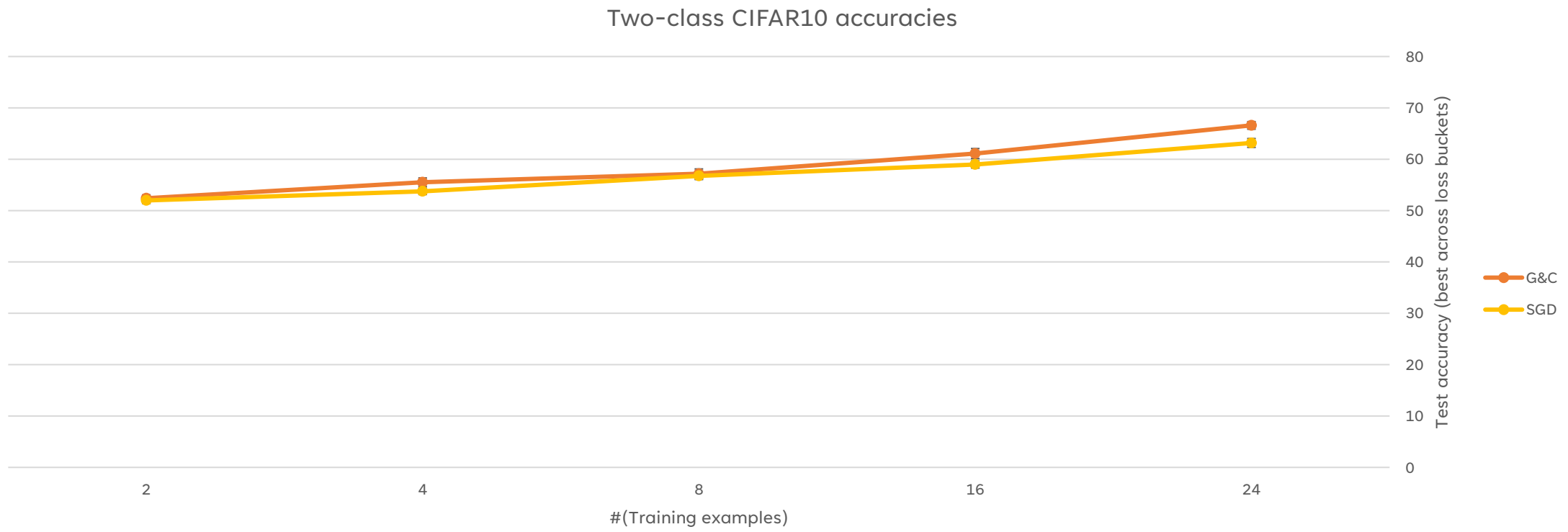
THE VOLUME HYPOTHESIS



AT SMALL SCALES, THIS IS DIRECTLY TESTABLE

- Guess & Check from $[-1,1]^N$ until the training data are fit (100% accuracy and to some loss threshold)
- Only bias is volume
- Work can scale exponentially with #(classes) and #(examples)

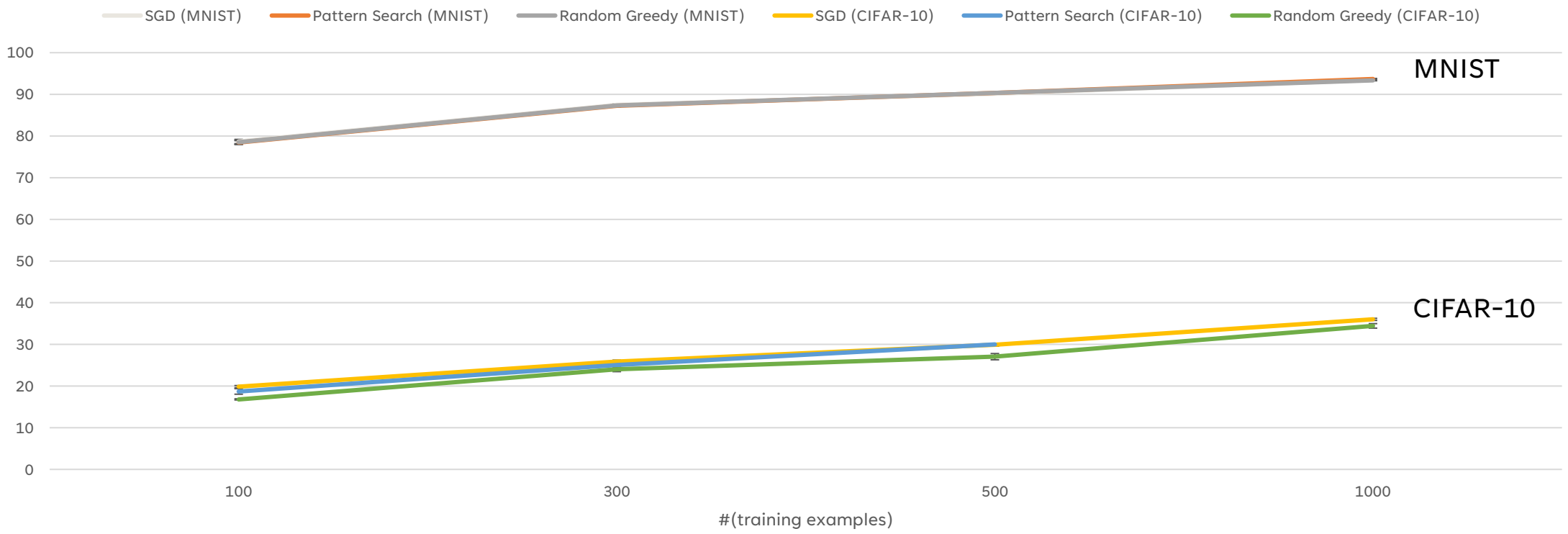
GUESS & CHECK GENERALIZES COMPARABLY



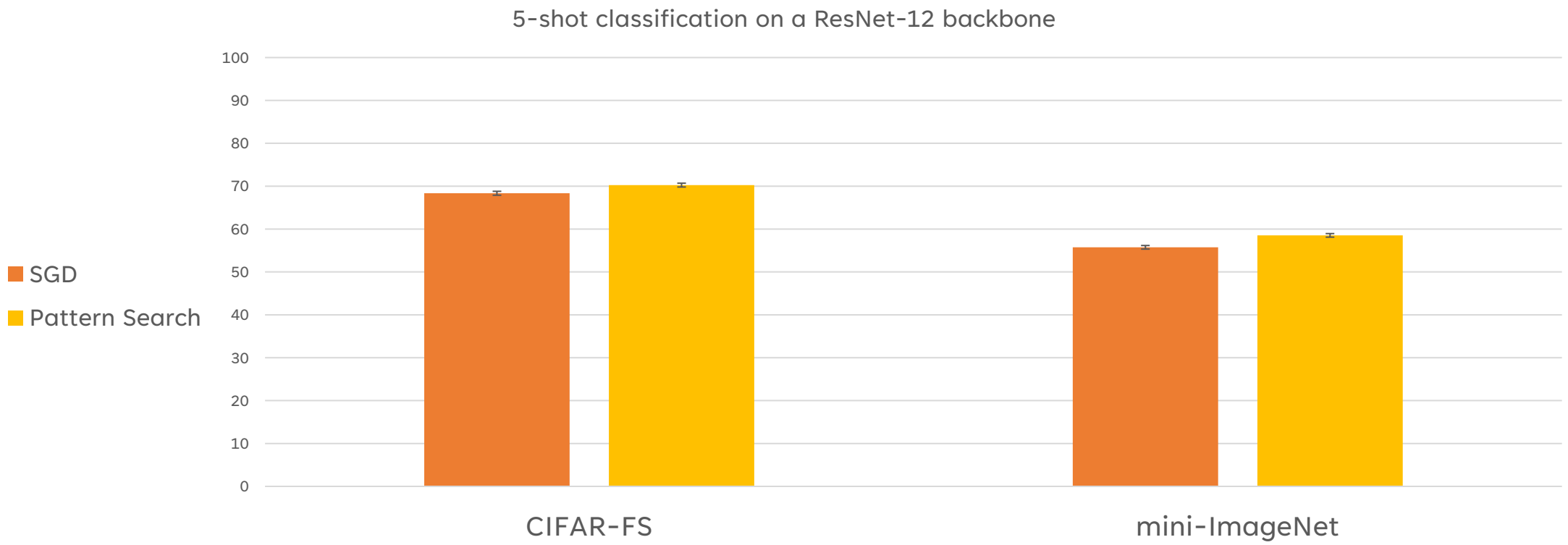
AND (KIND OF) TESTABLE AT LARGER SCALES

- Pattern search – pick random parameter, move it a fixed step, decrease step if no parameter works
- “Random greedy search” – add Gaussian noise, update the iterate if this decreases the loss

THESE ALSO GENERALIZE COMPARABLY TO SGD: MNIST/CIFAR-10, LARGER SAMPLES (PS, RG)



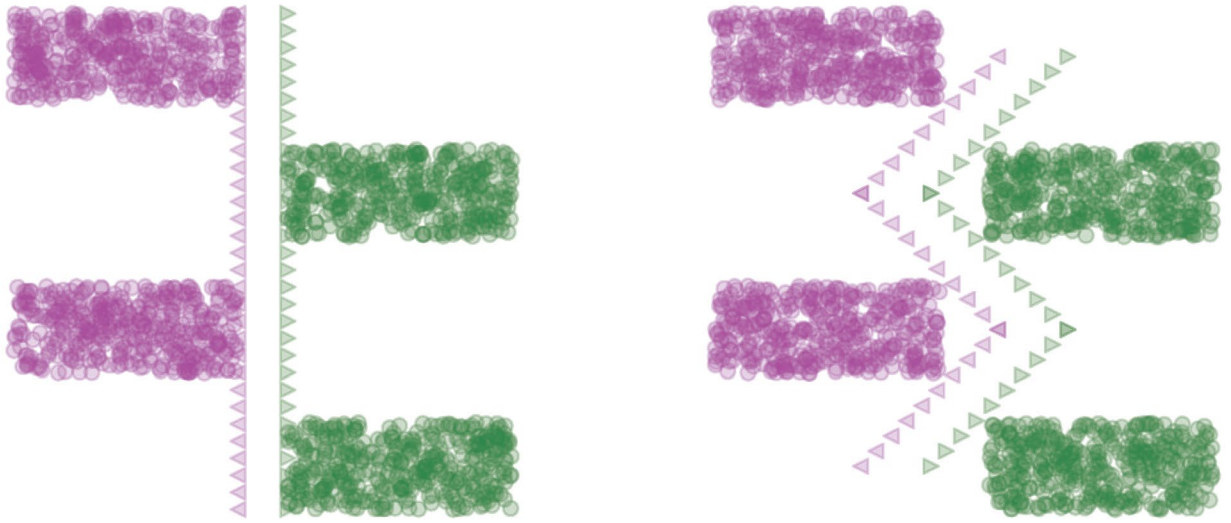
THESE ALSO GENERALIZE COMPARABLY TO SGD: FEW-SHOT LEARNING WITH PATTERN SEARCH



LIMITATIONS

- Only highly-overparametrized regime
- Smaller scale classification experiment
- Only Guess & Check is exclusively biased toward volume, other two could share behavior with GD
- No direct link shown between GD and volume

SIMPLICITY AND OTHER IMPLICIT BIASES?



Shah et al. “The Pitfalls of Simplicity Bias in Neural Networks”, NeurIPS’20.

Volume (based on guesses until success)	10^{-4}	$< 10^{-10}$

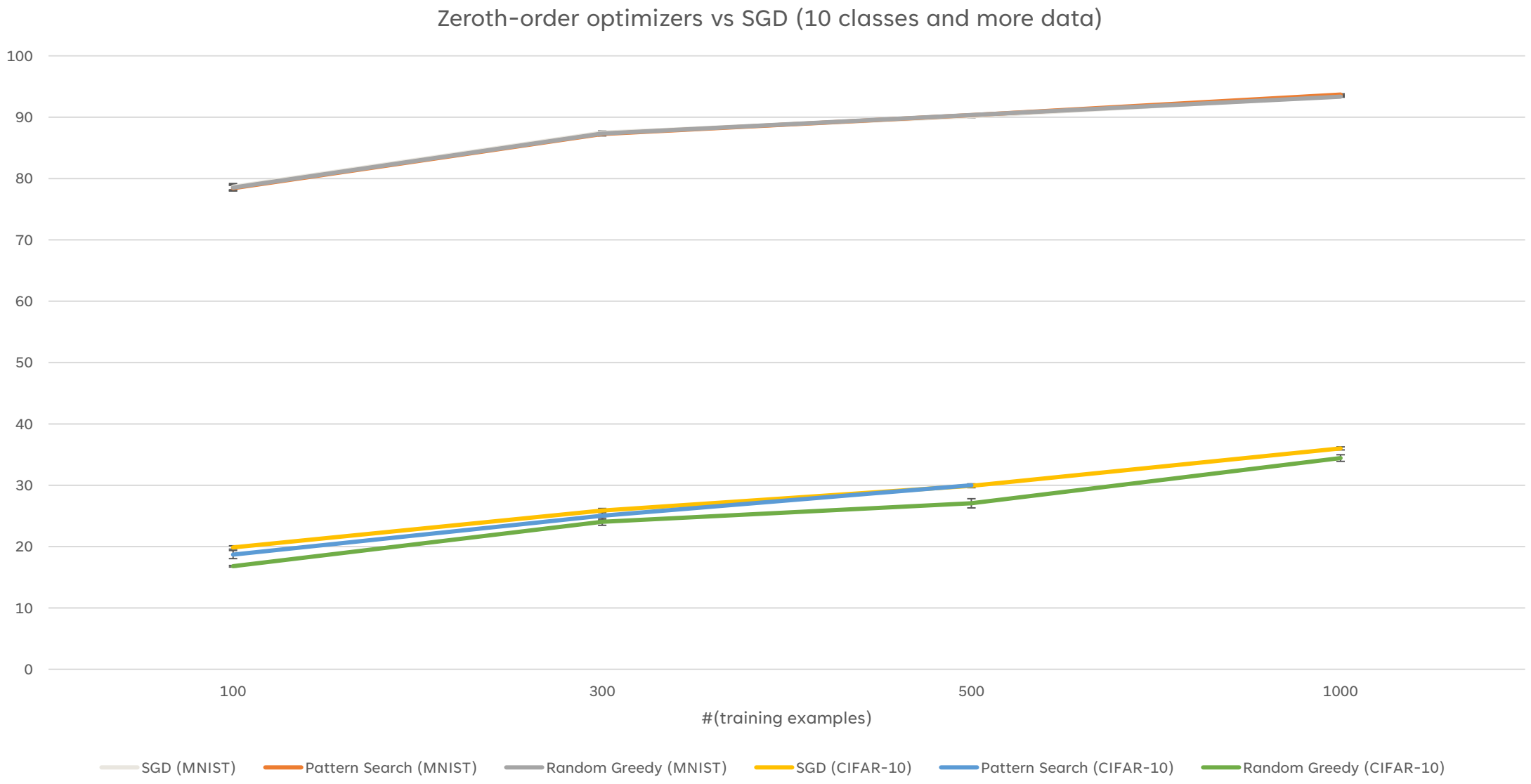
https://iclr.cc/virtual/2023/oral/12746

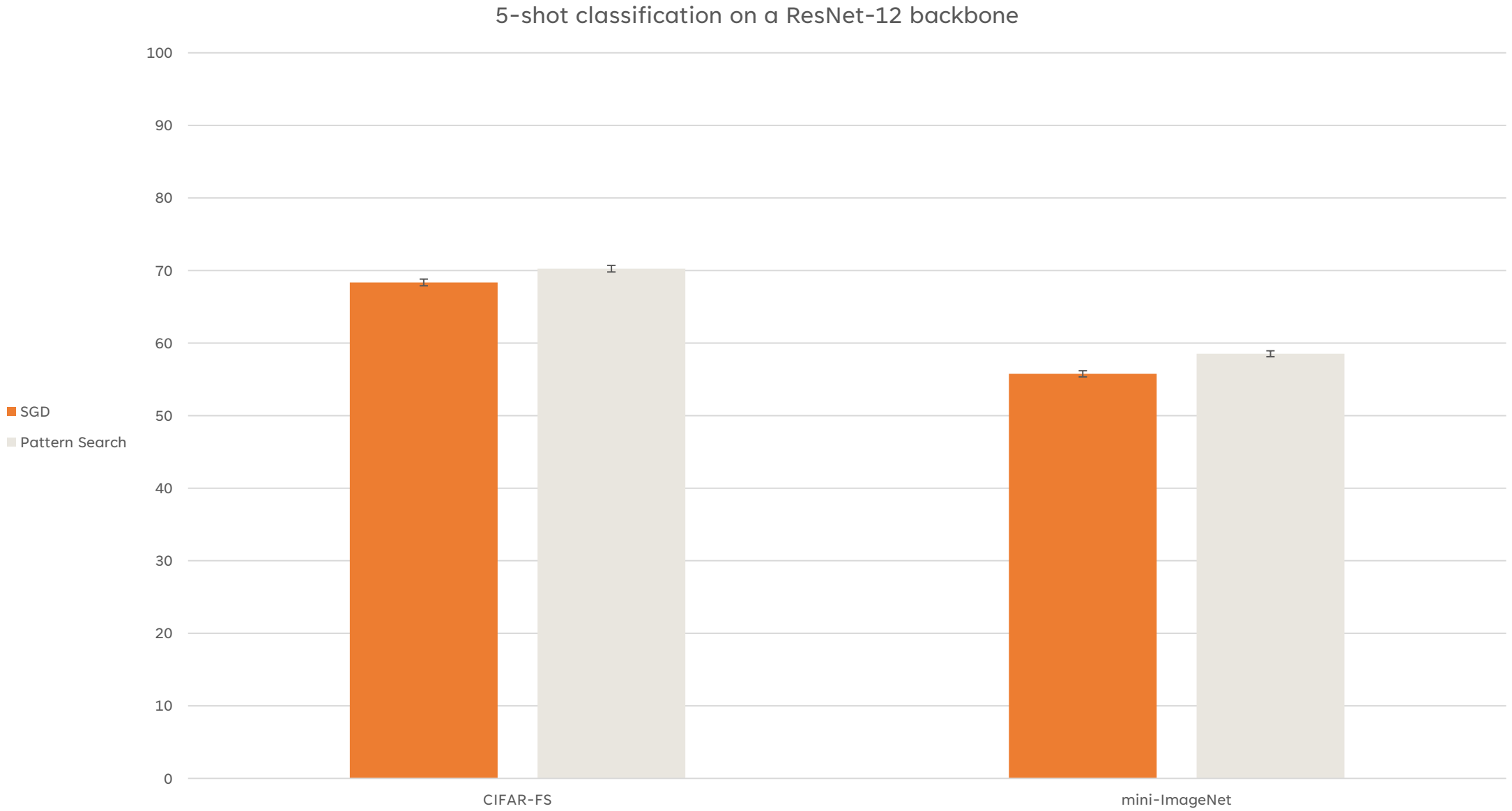
Poster #87

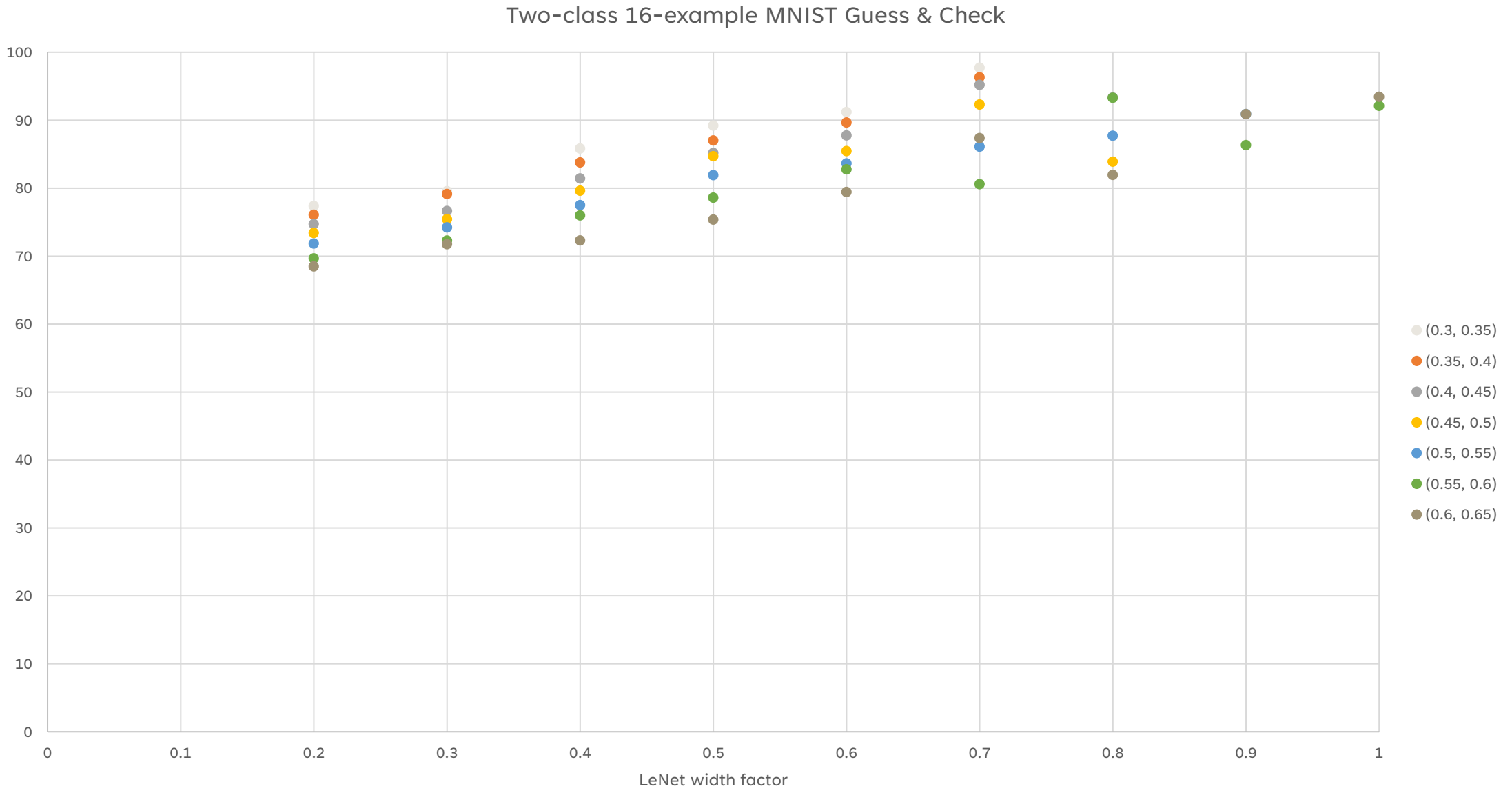
FIN

Not included here:

- G&C test performance scales with width
- Experimental variations, e.g., the sampling range doesn't matter - tried up to [-5,5]


Zeroth-order optimizers vs SGD (10 classes and more data)
100
90
80
70
60
50
40
30
20
10
0
100
300
500
1000
#(training examples)
SGD (MNIST)
Pattern Search (MNIST)
Random Greedy (MNIST)
SGD (CIFAR-10)
Pattern Search (CIFAR-10)
Random Greedy (CIFAR-10)


5-shot classification on a ResNet-12 backbone
100
90
80
70
60
50
40
30
20
10
0
SGD
Pattern Search
CIFAR-FS
mini-ImageNet


Two-class 16-example MNIST Guess & Check
100
90
80
70
60
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
LeNet width factor
(0.3, 0.35)
(0.35, 0.4)
(0.4, 0.45)
(0.45, 0.5)
(0.5, 0.55)
(0.55, 0.6)
(0.6, 0.65)