ICLR 2021

Are wider nets better given the same number of parameters?

github.com/google-research/wide-sparse-nets

Anna Golubeva[1*]
agolubeva@pitp.ca

Behnam Neyshabur[2]
neyshabur@google.com

Guy Gur-Ari[2]
guyga@google.com

[1]Perimeter Institute for Theoretical Physics, Waterloo, ON, Canada

[*]Work done while an intern at Blueshift.

[2]Blueshift, Alphabet, Mountain View, CA

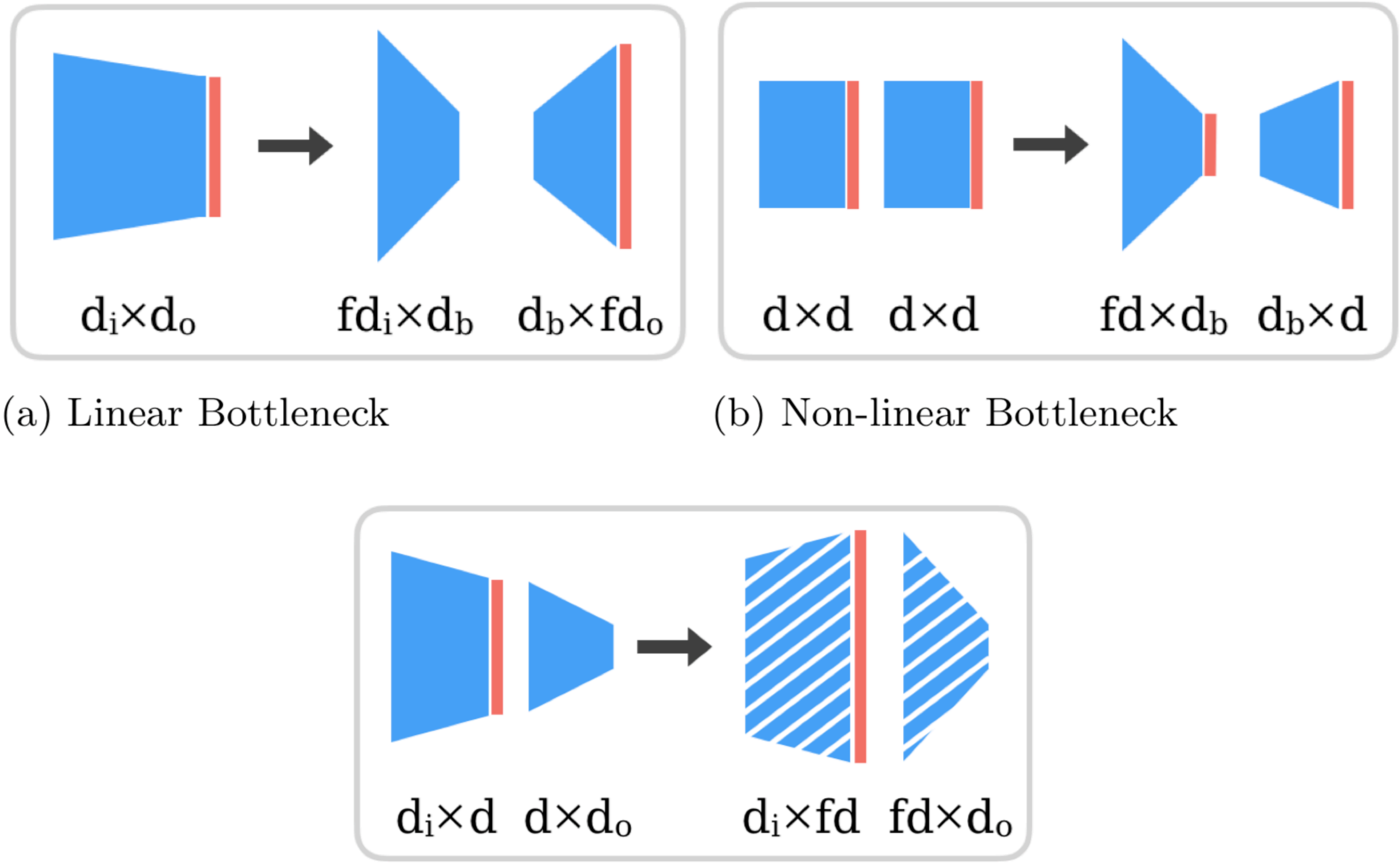
- Increasing the number of NN parameters improves performance.

- Increasing the number of NN parameters improves performance.
- The number of parameters is increased along with layer width.

- Increasing the number of NN parameters improves performance.
- The number of parameters is increased along with layer width.

► Is the performance gain due to
more params or larger width?

How to increase width independently from the number of params?

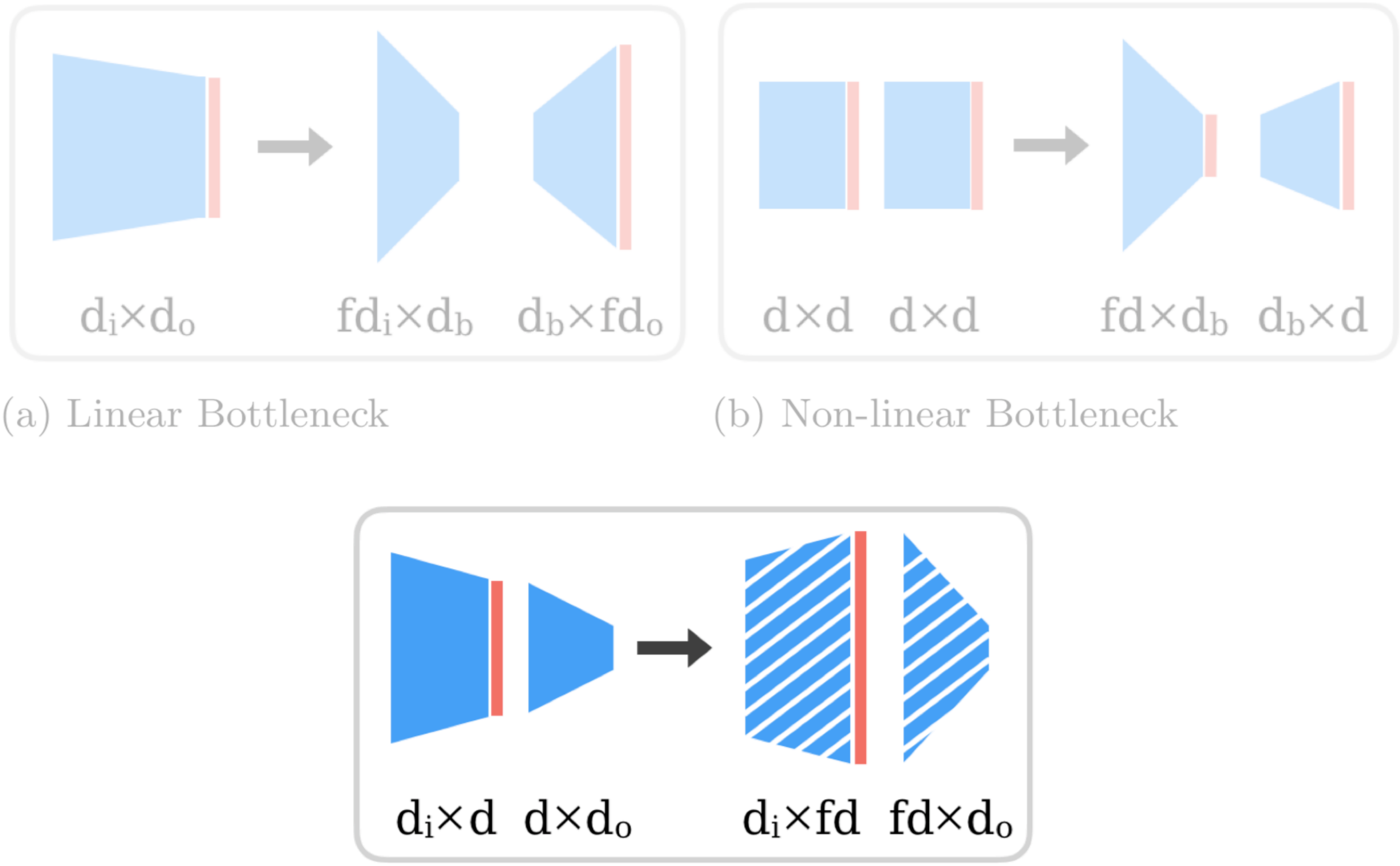


(a) Linear Bottleneck

(b) Non-linear Bottleneck

(c) Static Sparsity

How to increase width independently from the number of params?

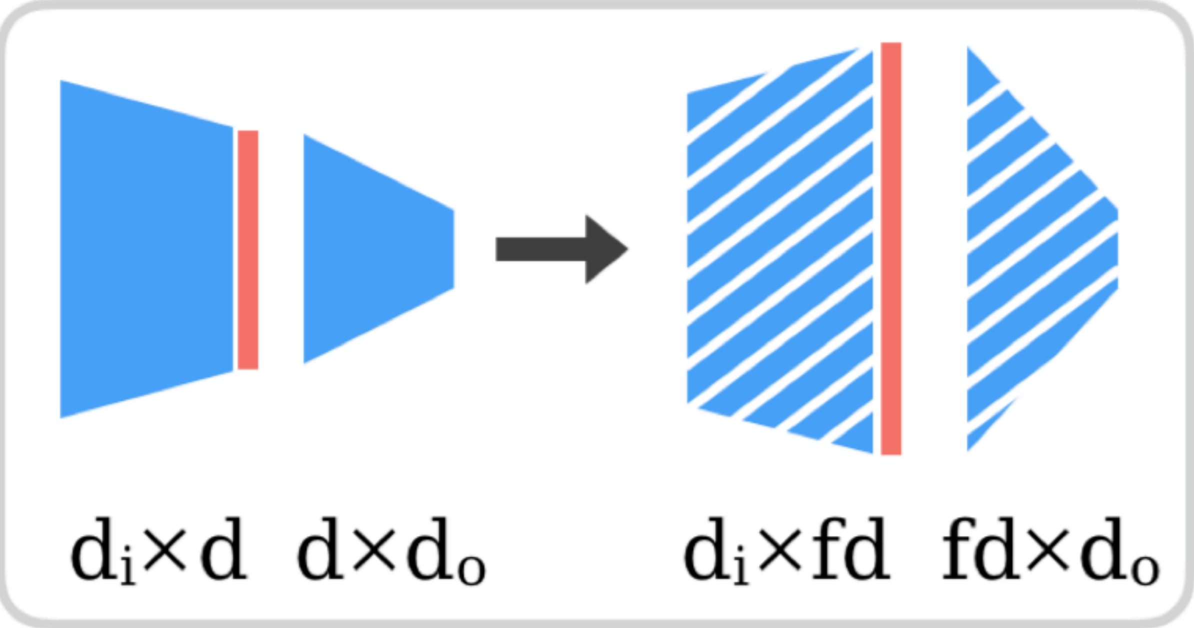


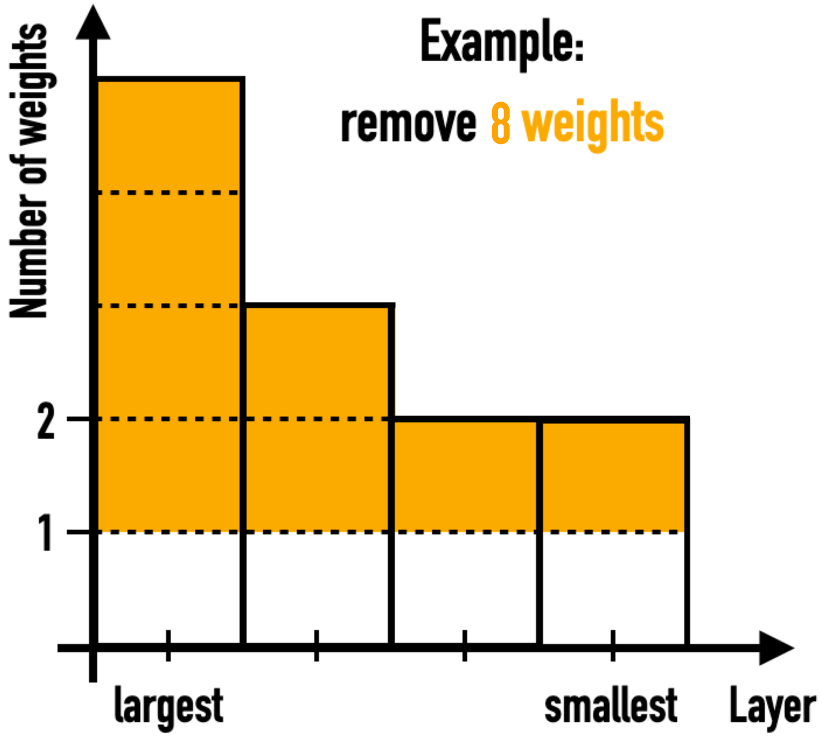
(a) Linear Bottleneck

(b) Non-linear Bottleneck

(c) Static Sparsity

Static Sparsity





- sparsity pattern: random, applied at initialization, static
- in-layer distribution uniform across all layer dimensions
- per-layer distribution according to layer size

- method advantage: it does not alter the NN architecture

! we are not aiming for performance gains through sparsity

Our approach in summary:

- select model type and architecture
- fix the number of weights

e.g. ResNet18 with
8 output channels
in the first conv layer

baseline: dense model (full connectivity)

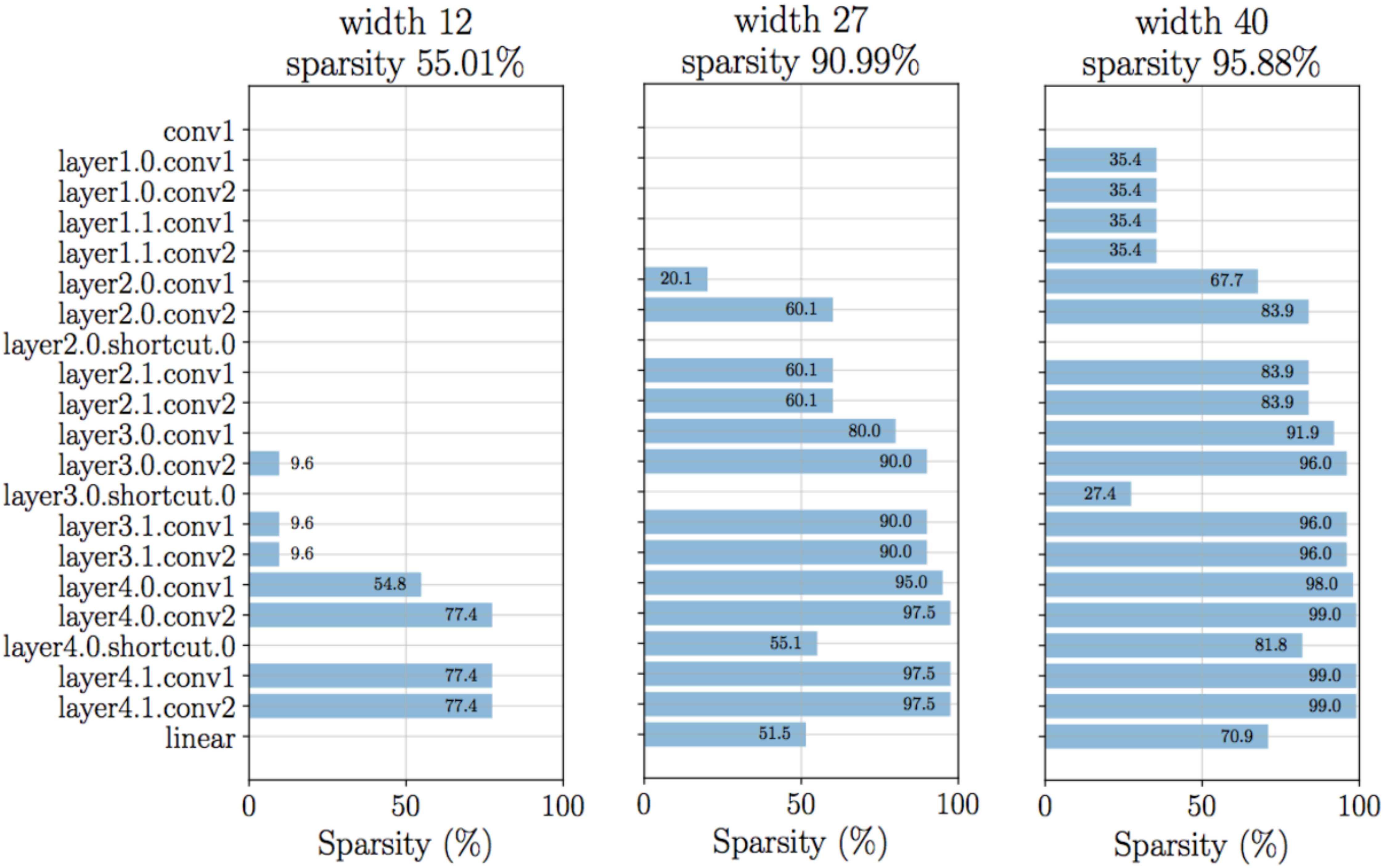
- build a family of models having different widths and sparsity levels but same number of weights

wide & sparse: increase the width and
remove excess weights

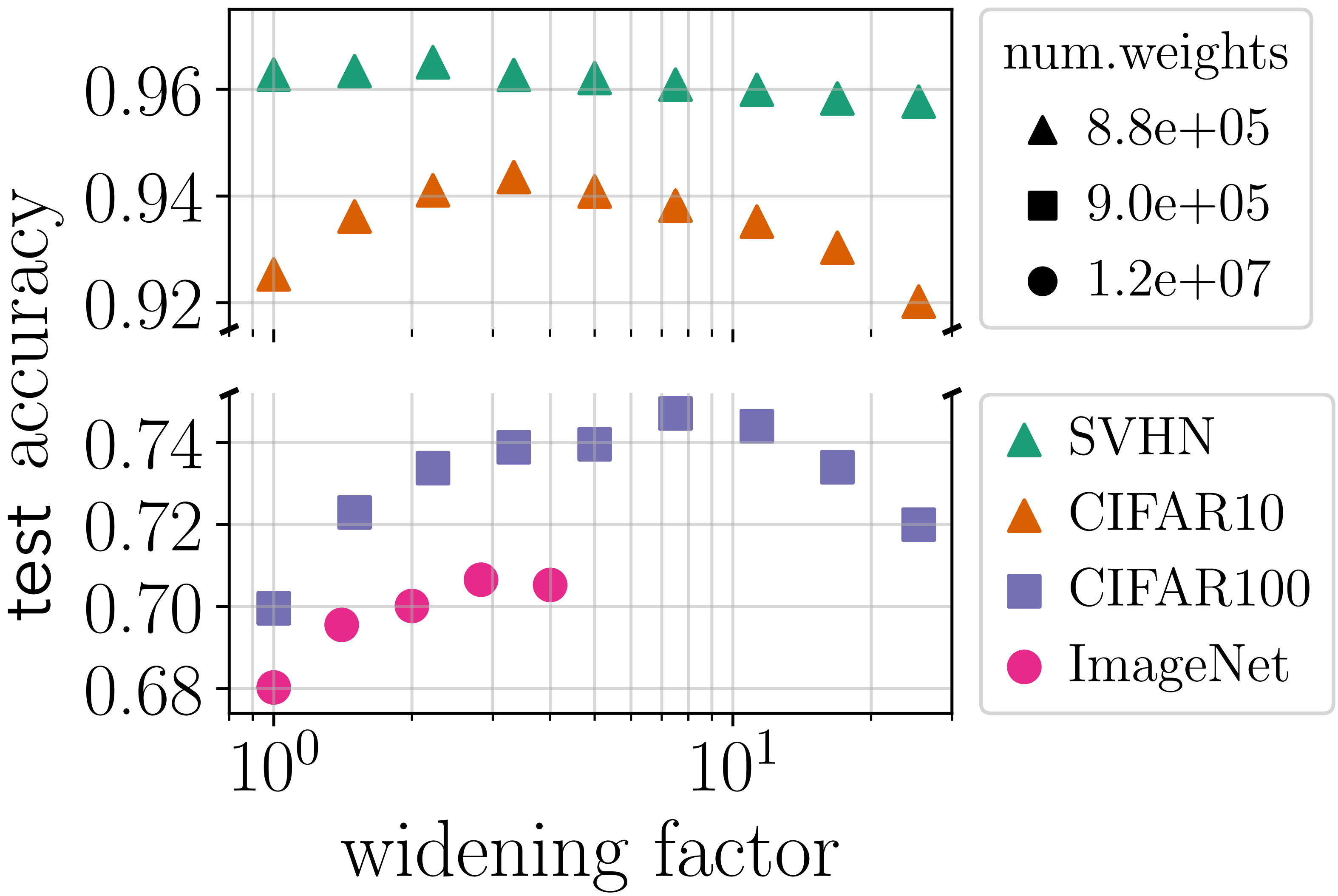
- train and compare performance

(task: image classification)

Sparsity distribution in a ResNet18 with base width 8



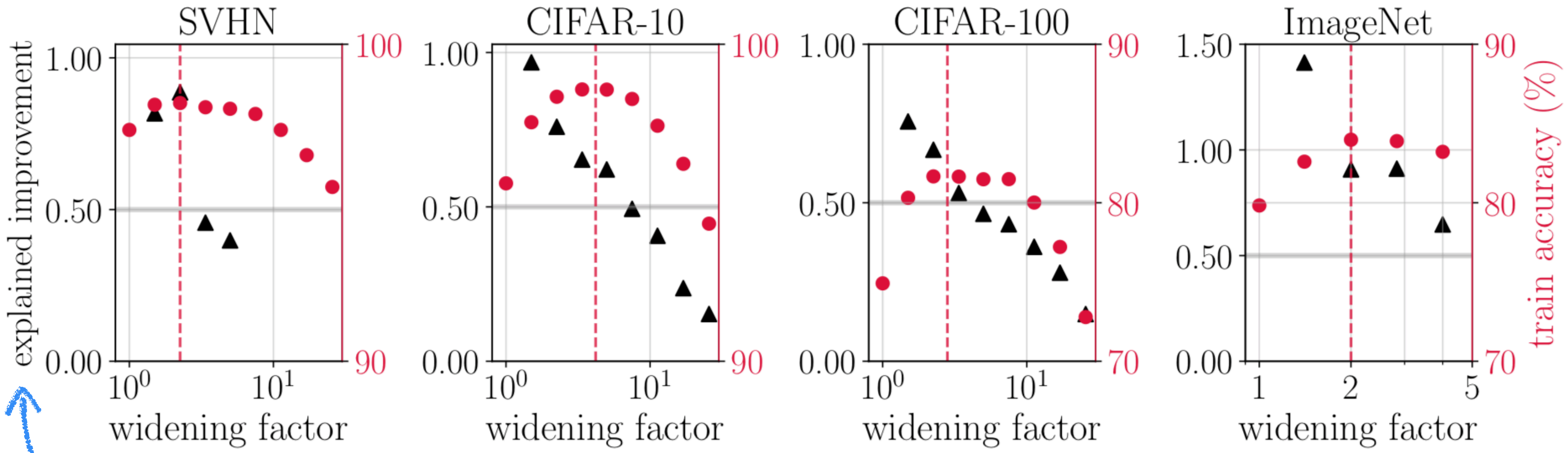
Results: ResNet-18



- test accuracy increases with the width, even though the number of weights is fixed

How much improvement is due to width only?

- compare perf increase for wide & sparse to wide & dense models:



$$= \frac{\Delta \text{ test acc. (sparse)}}{\Delta \text{ test acc. (dense)}}$$

► as long as the model can achieve high training accuracy,
most of the improvement in performance can be attributed to the width

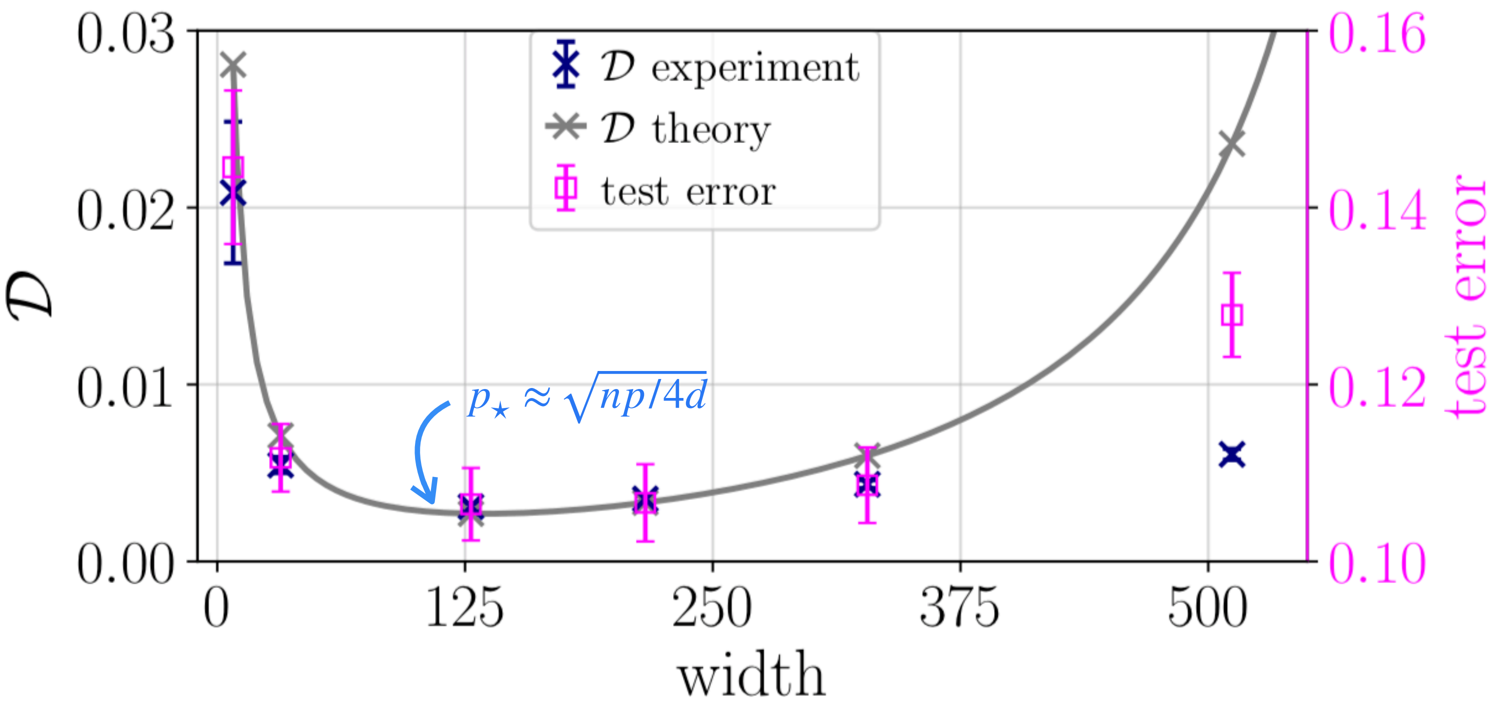
Theory: ∞-width limit and GP kernels

Gaussian Process

- hypothesis: performance improvement is correlated with having a GP kernel that is closer to the ∞-width kernel
- hypothesis: the distance to the ∞-width kernel can be reduced by increasing network width

► perf. correlates strongly with the distance to the ∞width kernel:

distance to ∞width kernel



theory predicts optimal connectivity

$$p_\star \approx \sqrt{\frac{np}{4d}}$$

with np= const.

input dimension d

layer width n