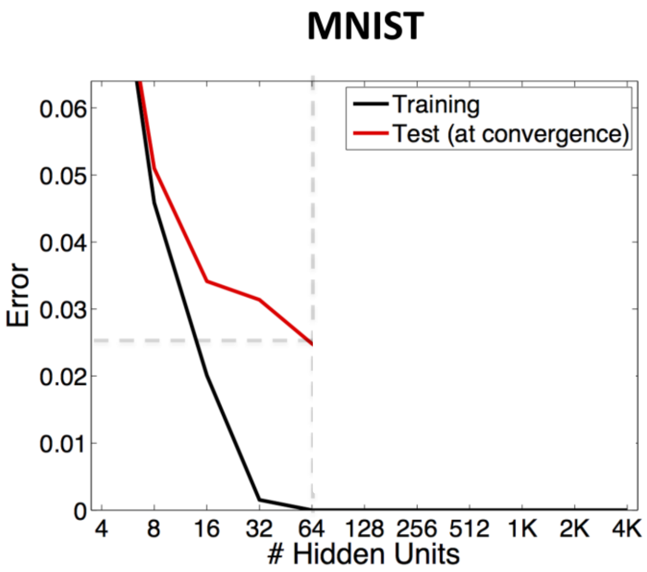
CPSC 540: Machine Learning

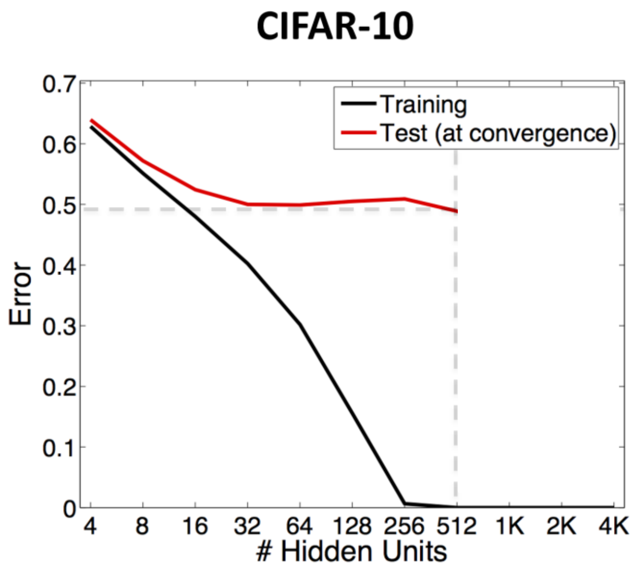
Double Descent Curves

Winter 2020

“Hidden” Regularization in Neural Networks

- Fitting single-layer neural network with SGD and no regularization:



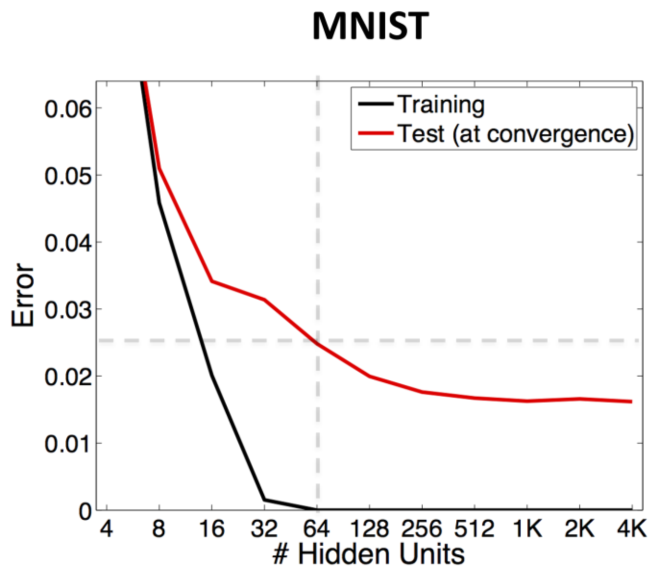


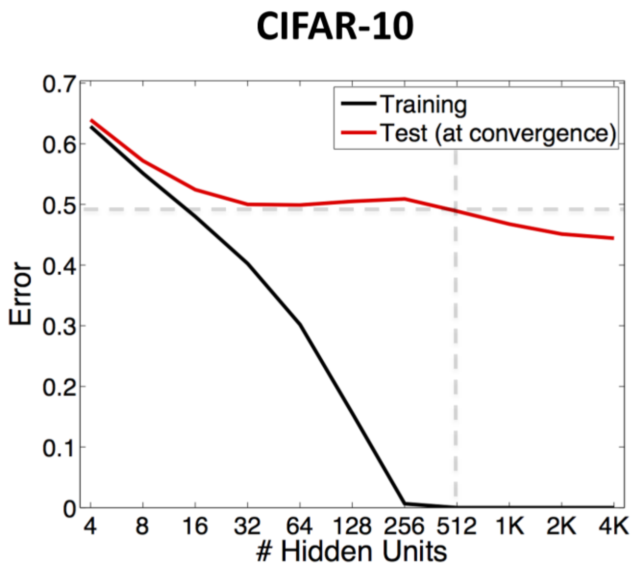
- Training goes to 0 with enough units: we’re finding a global min.

- What should happen to training and test error for larger #hidden?

“Hidden” Regularization in Neural Networks

- Fitting single-layer neural network with SGD and no regularization:



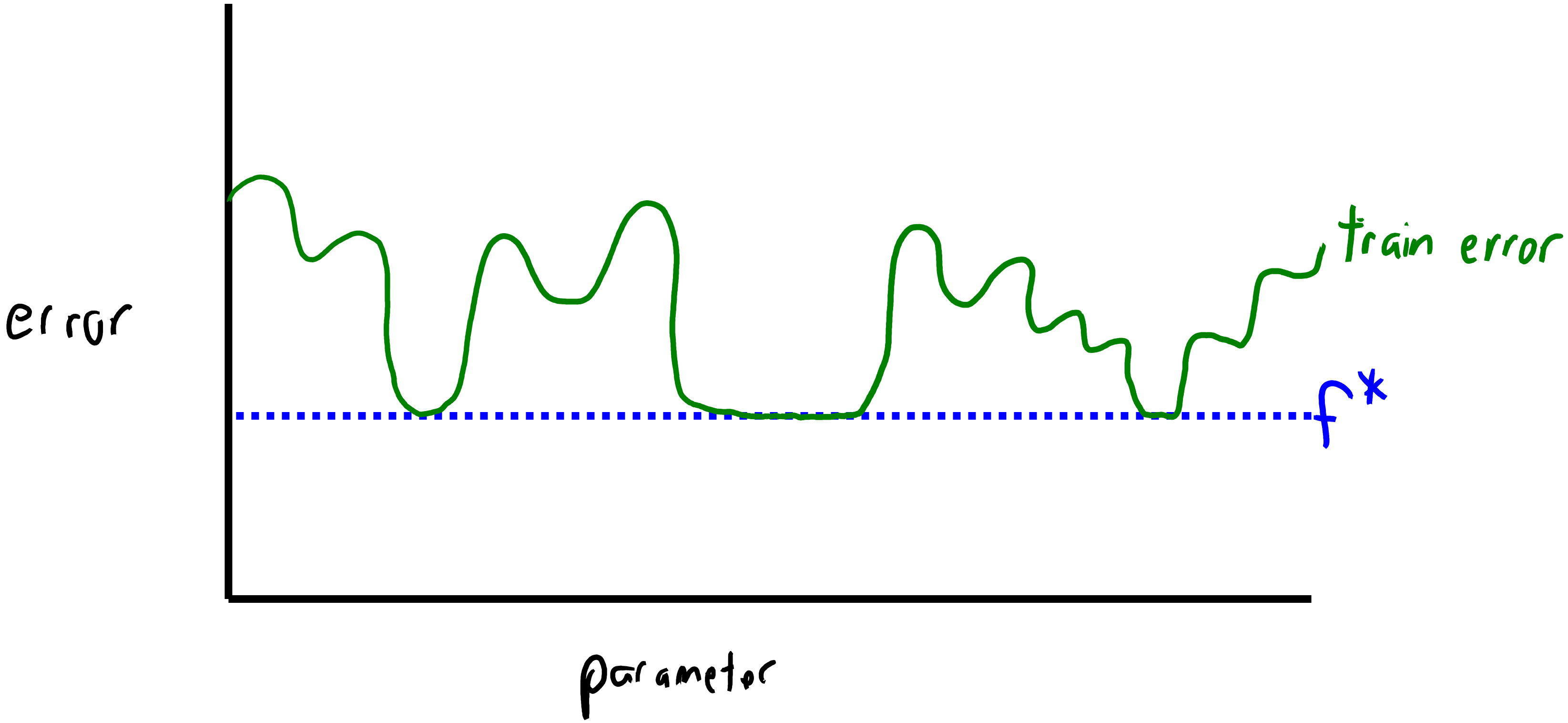


- Test error continues to go down!?! Where is fundamental trade-off??
- There exist global mins with large #hidden units have test error = 1.
  - But among the global minima, SGD is somehow converging to “good” ones.

https://www.neyshabur.net/papers/inductive_bias_poster.pdf

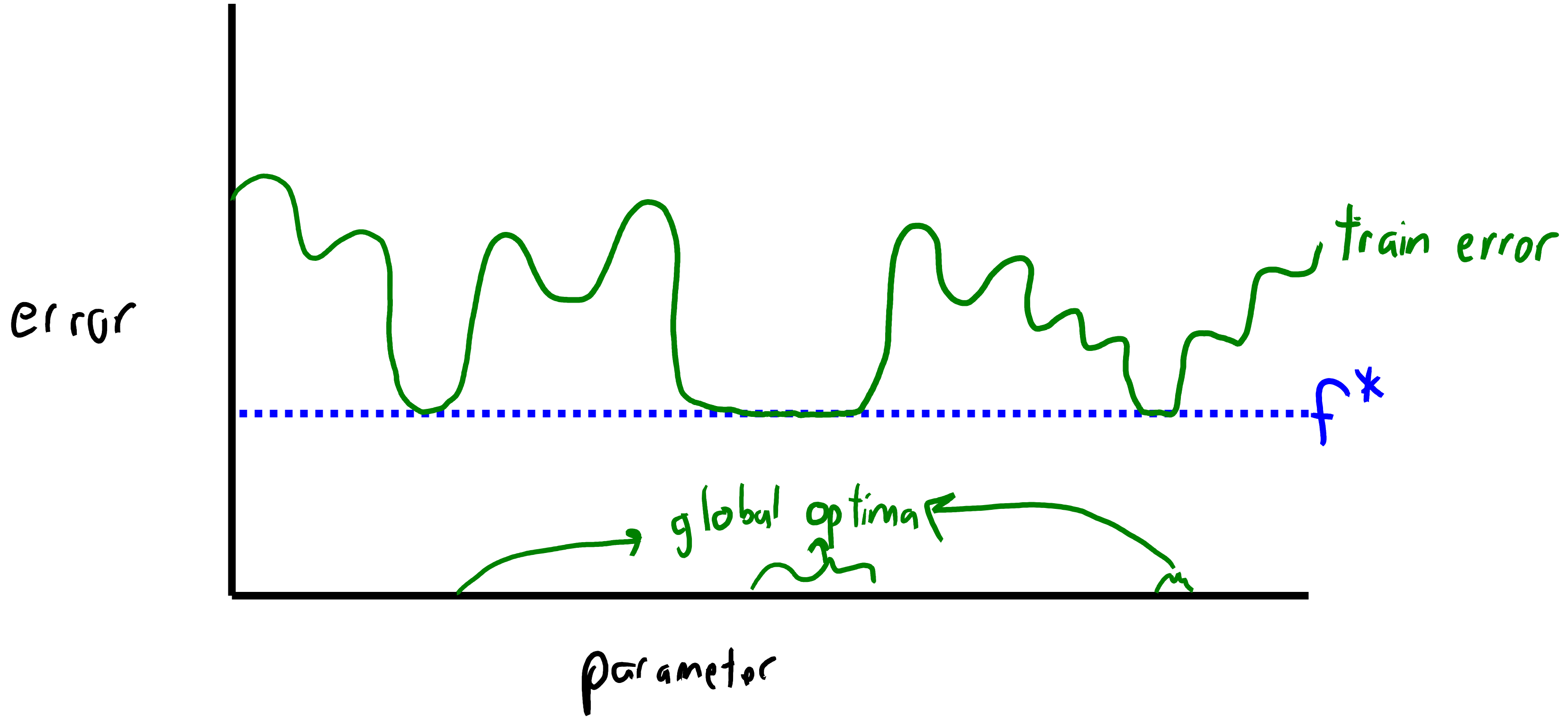
Multiple Global Minima?

- For standard objectives, there is a global min function value f*:



Multiple Global Minima?

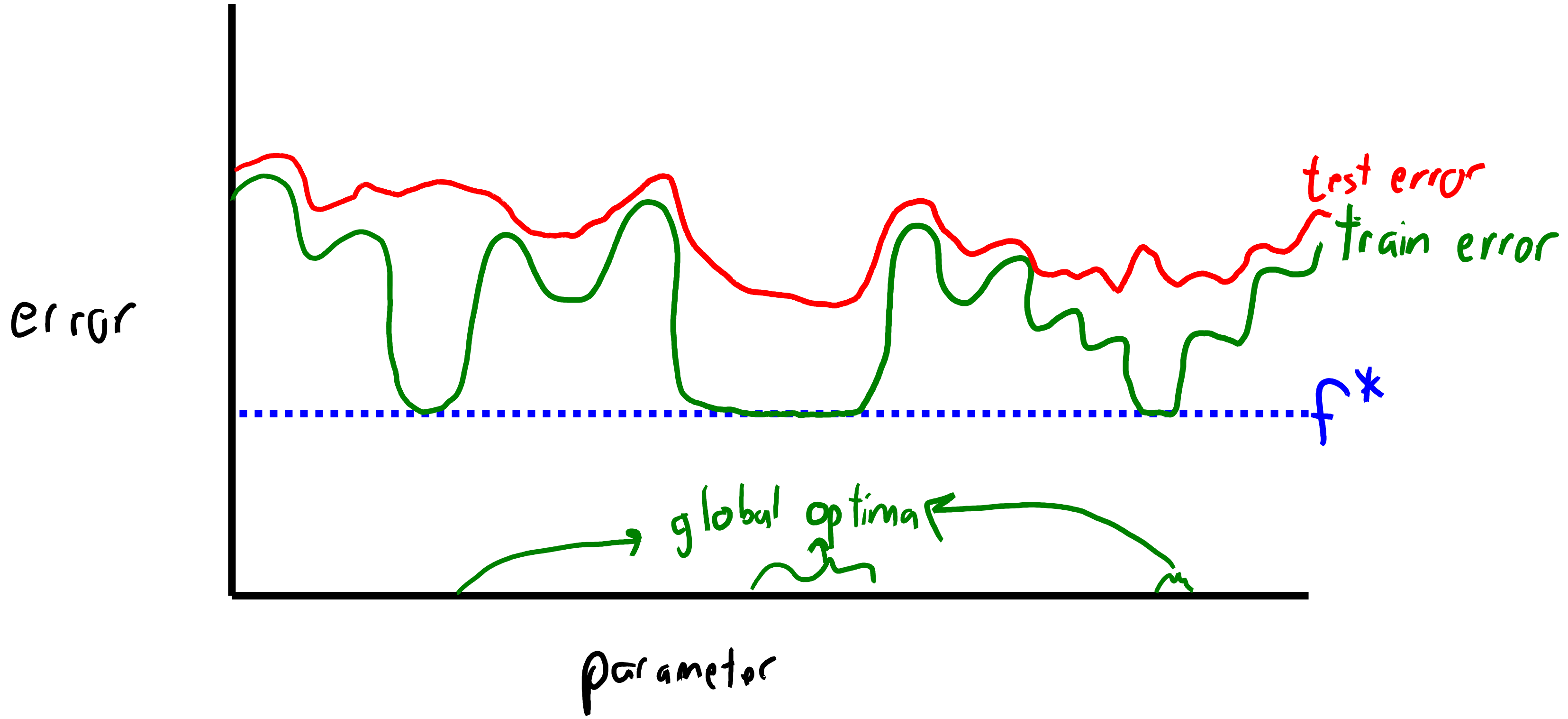
- For standard objectives, there is a global min function value f*:



- But this may be achieved by many different parameter values.

Multiple Global Minima?

- For standard objectives, there is a global min function value f*:



- But this may be achieved by many different parameter values.
  - These training error "global minima" may have very-different test errors.
  - Some of these global minima may be more "regularized" than others.

Implicit Regularization of SGD

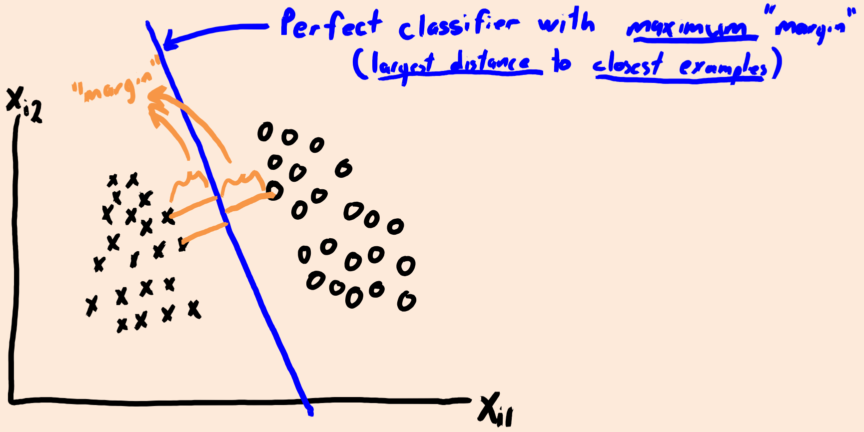
- There is growing evidence that using SGD regularizes parameters.
  - We call this the "implicit regularization" of the optimization algorithm.

- Beyond empirical evidence, we know this happens in simpler cases.

- Example of implicit regularization:
  - Consider a least squares problem where there exists a 'w' where Xw=y.
    - Residuals are all zero, we fit the data exactly.
  - You run [stochastic] gradient descent starting from w=0.
  - Converges to solution Xw=y that has the minimum L2-norm.
    - So using SGD is equivalent to L2-regularization here, but regularization is "implicit".

Implicit Regularization of SGD

- Example of implicit regularization:
  - Consider a logistic regression problem where data is linearly separable.
    - We can fit the data exactly.
  - You run gradient descent from any starting point.
  - Converges to max-margin solution of the problem.
    - So using gradient descent is equivalent to encouraging large margin.



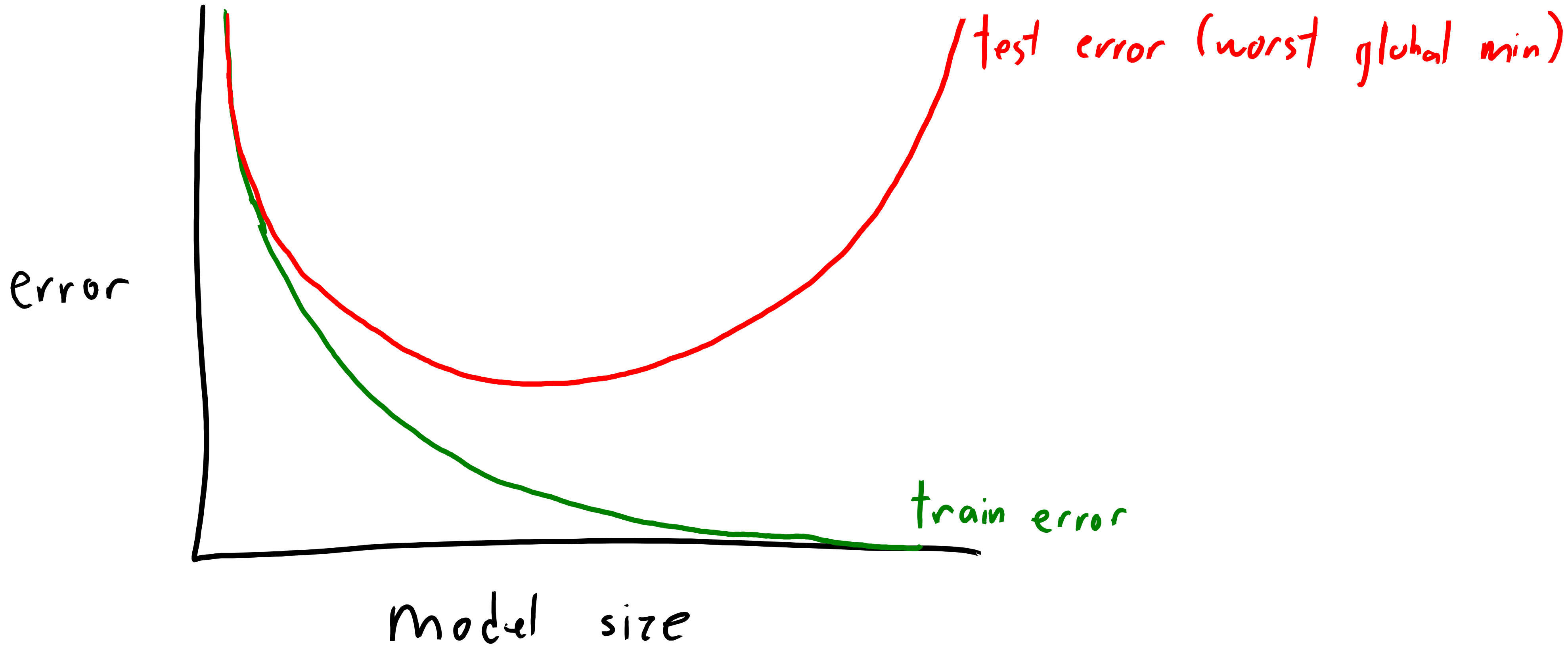
- Similar result known for boosting and matrix factorization.

Double Descent Curves

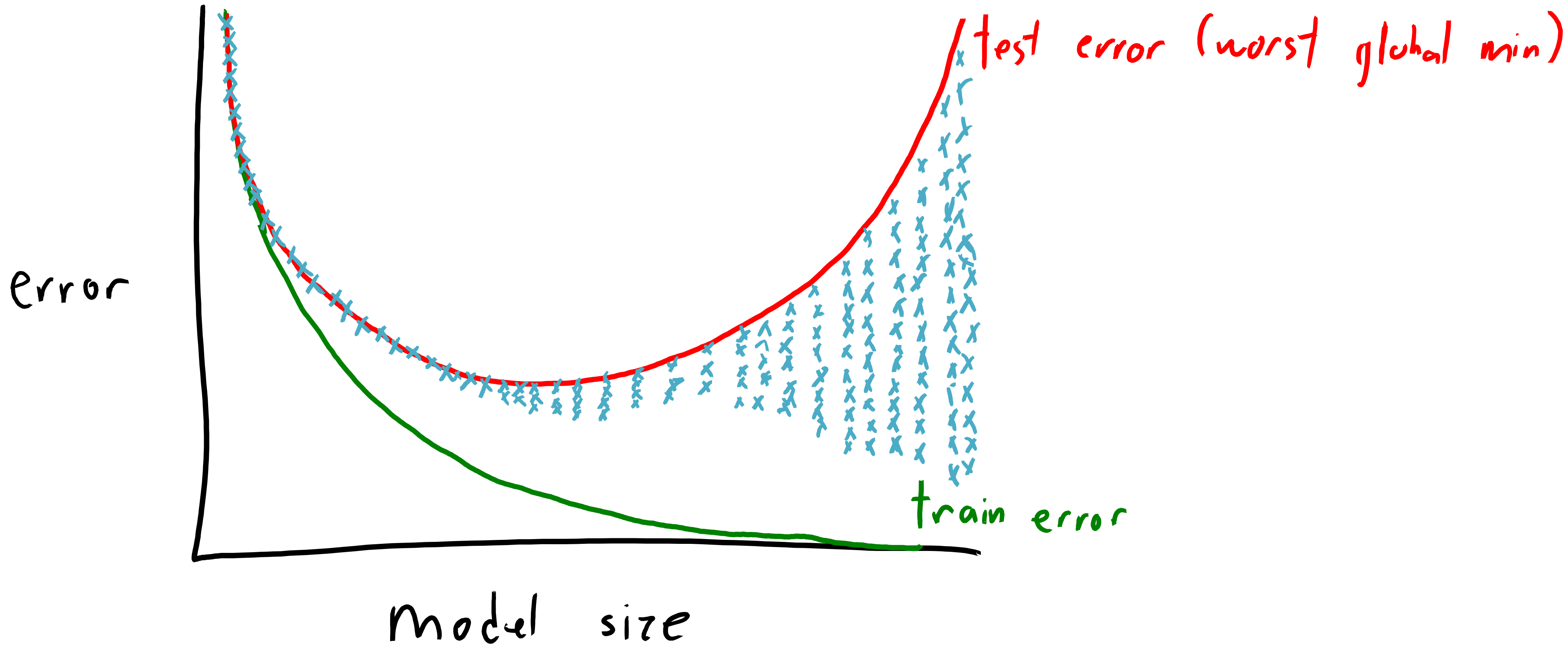
- What is going on???

https://openai.com/blog/deep-double-descent/

Worst vs. Best "Global Minimum"

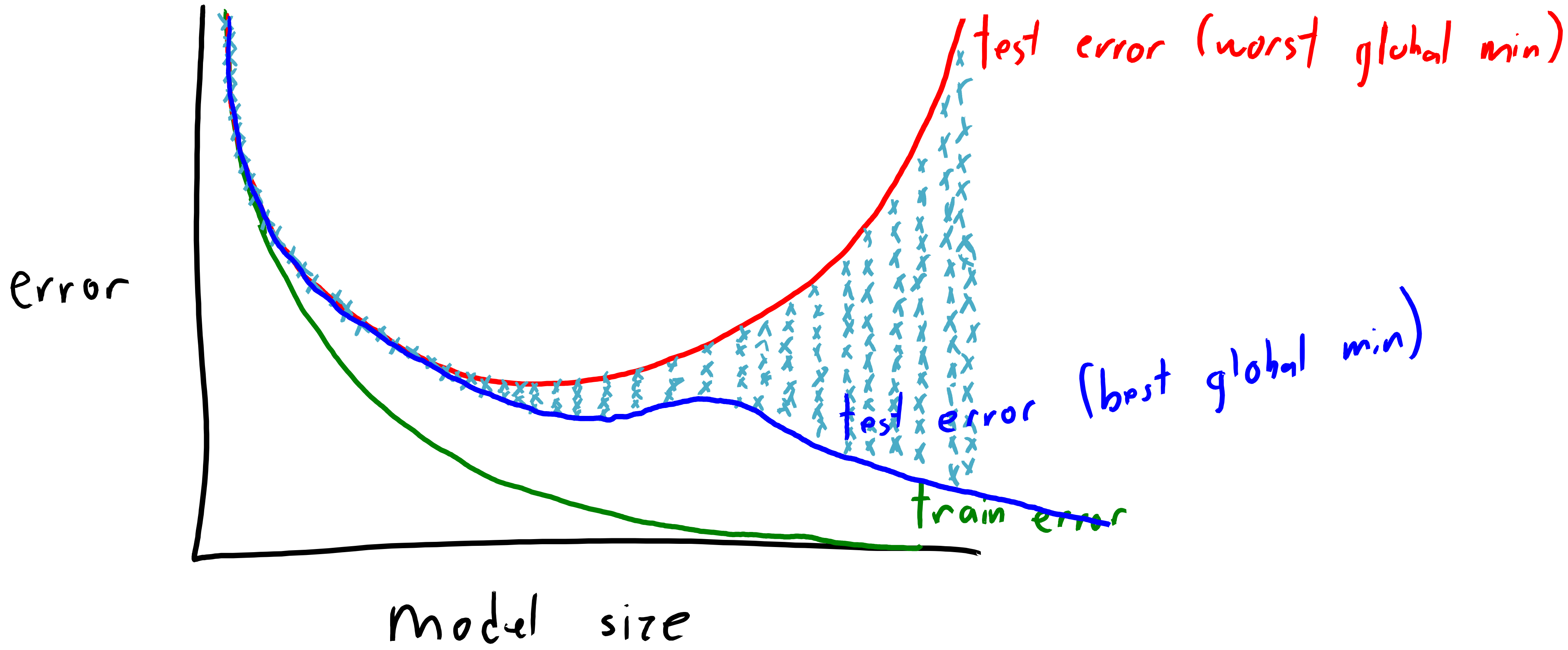


Worst vs. Best "Global Minimum"



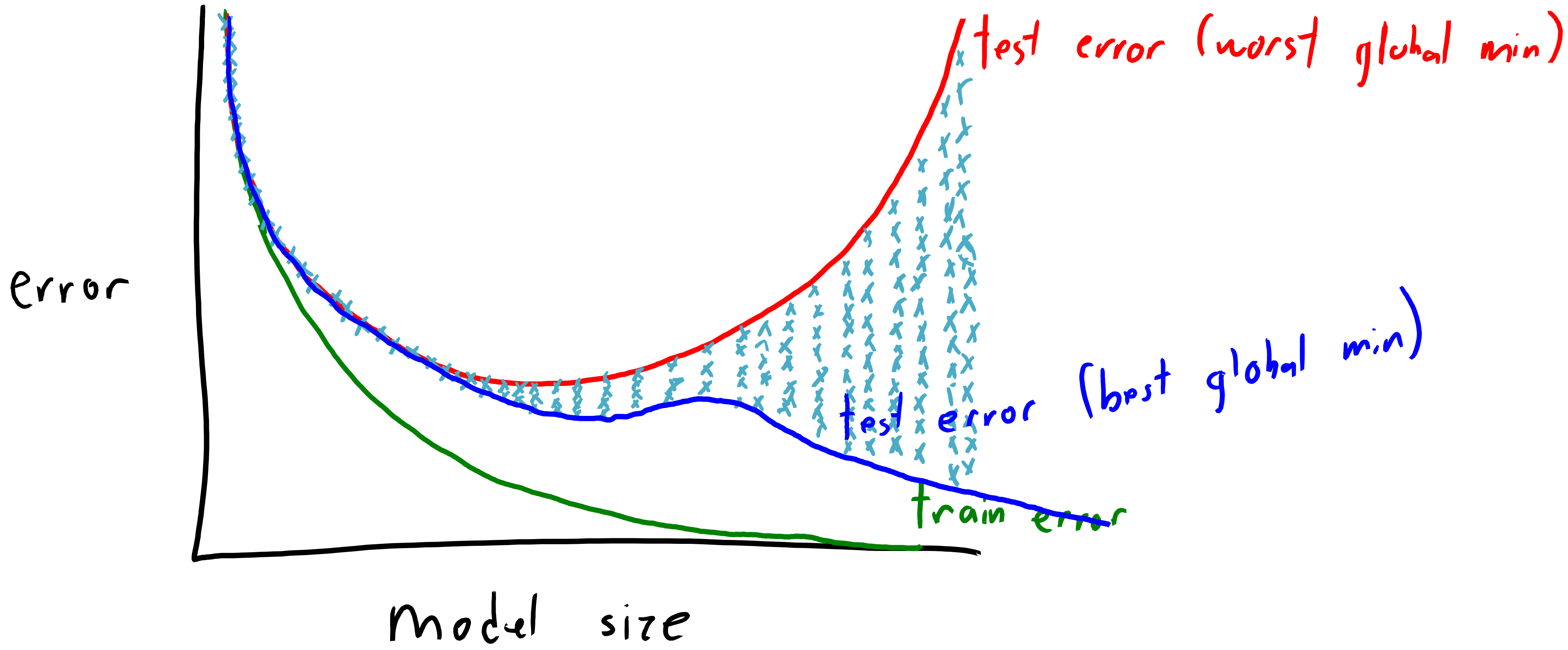
- Learning theory results analyze global min with worst test error.
  - Actual test error for different global minima be better than worst case bound.
  - Theory is correct, but maybe "worst overfitting possible" is too pessimistic?

Worst vs. Best "Global Minimum"



- Consider instead the global min with best test error.
  - With small models, "minimize training error" leads to unique (or similar) global mins.
  - With larger models, there is a lot of flexibility in the space of global mins (gap between best/worst).
- Gap between "worst" and "best" global min can grow with model complexity.

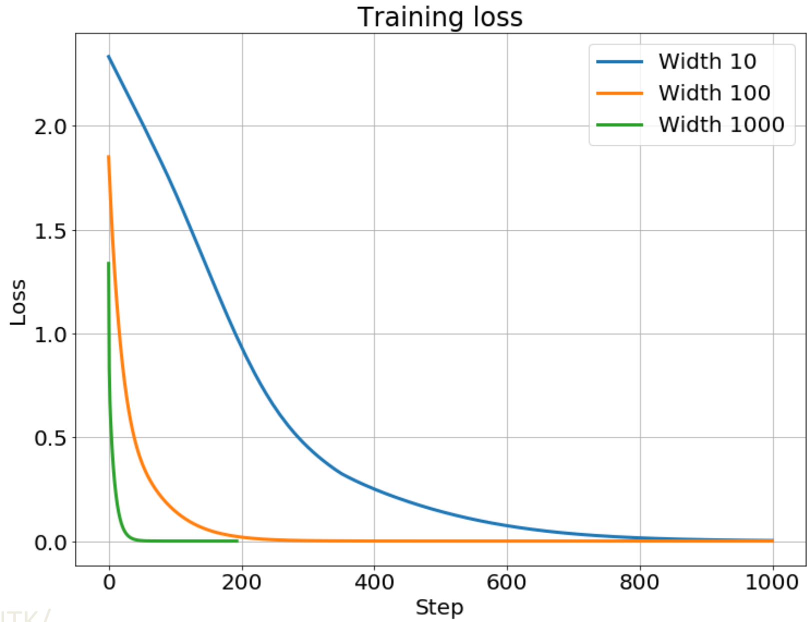
Worst vs. Best "Global Minimum"

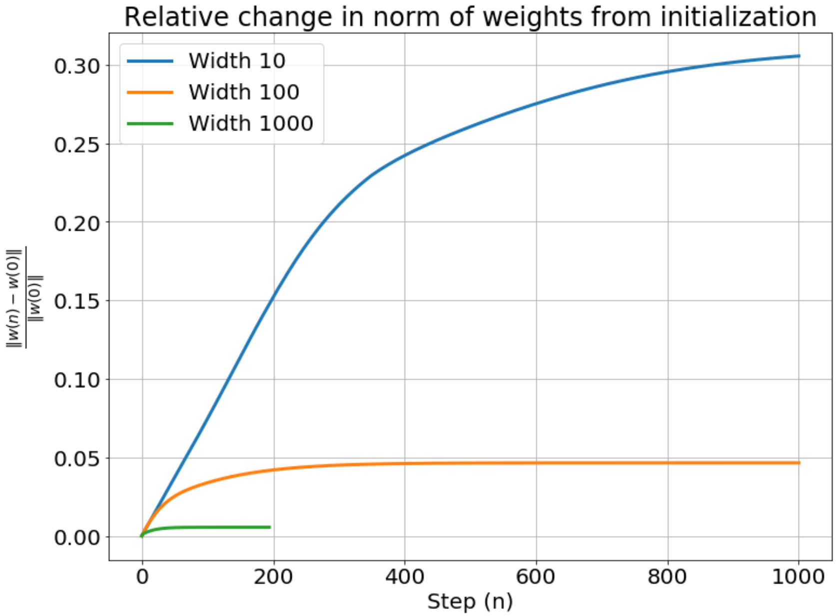


- Can get "double descent" curve in practice if parameters roughly track "best" global min shape.
  - One way to do this: increase regularization as you increase model size.
- Maybe "neural network trained with SGD" has "more implicit regularization for bigger models"?
  - But this behavior is not specific to implicit regularization of SGD and not specific to neural networks.

Implicit Regularization of SGD (as function of size)

- Why would implicit regularization of SGD increase with dimension?
  - Maybe SGD finds low-norm solutions?
    - In higher-dimensions, there is flexibility in global mins to have a low norm?
  - Maybe SGD stays closer to starting point as we increase dimension?
    - This would be more like a regularizer of the form $||w - w^0||$.





https://rajatvd.github.io/NTK/

Summary

- Neural networks learn features for supervised learning.
  - For structured prediction, may reduce need to rely on inference.
- Implicit regularization and double descent curves.
  - Possible explanations for why deep networks often generalize well.
- Next time: combining deep learning with the rest of the course.