Regularization

Lyle Ungar

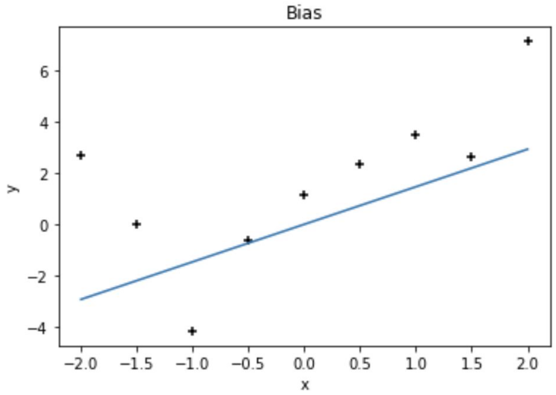
neuromatch academy

The goal of supervised learning is generalization

- **We minimize error on a training set**
  - using a ***really*** complex model
- **But we care about the error on a (future) test set**
  - want to fit the signal, not the noise
- **Why not use a simpler model?**
  - it doesn't work as well

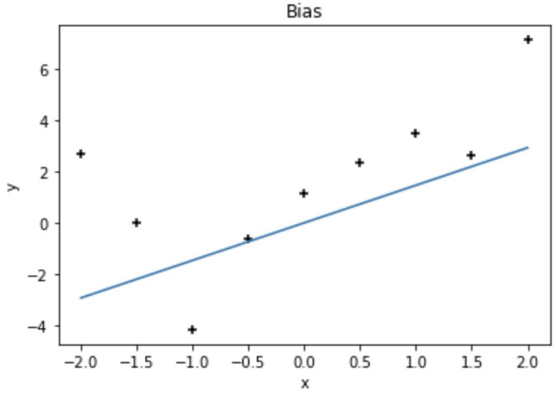
Model complexity matters

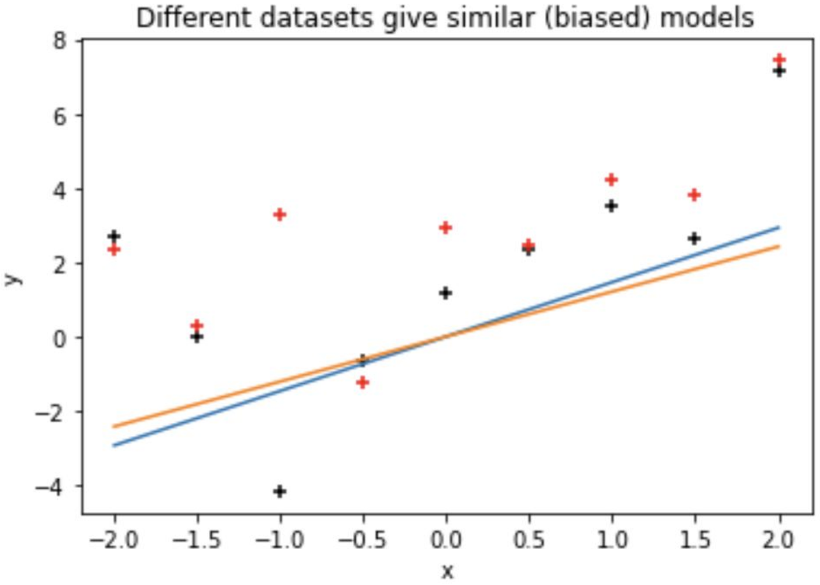
- Too simple a model “underfits”
  - It fails to capture the signal in the data



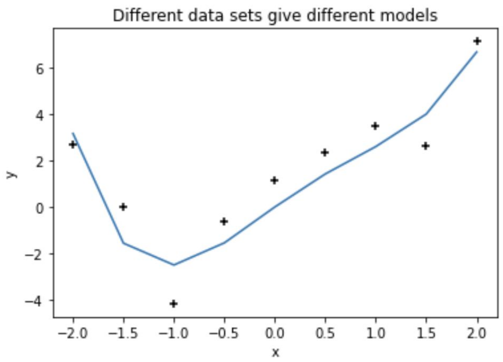
Bias

- Simple models can underfit
  - weights are systematically too small



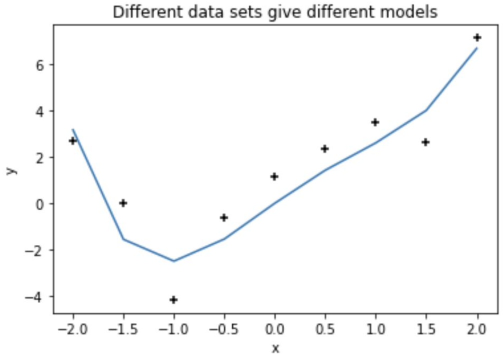


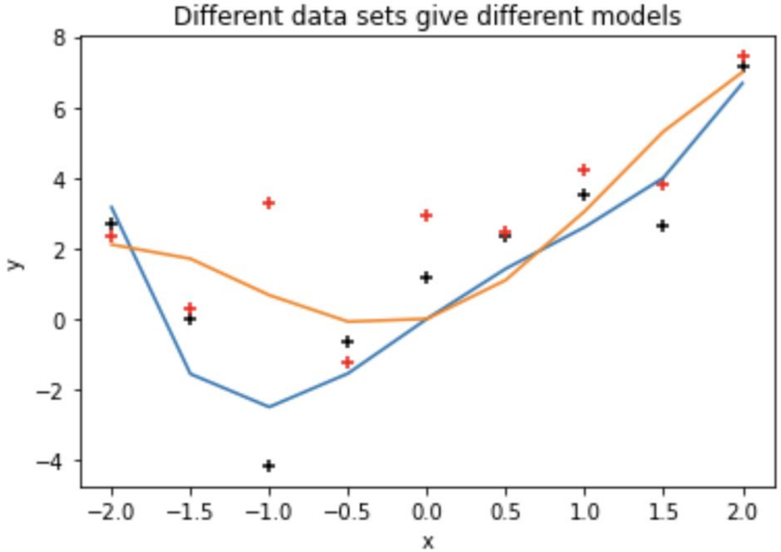
Model complexity matters

- Too complex a model "overfits"
    - It fits the noise in the data, and so generalizes poorly
    - 

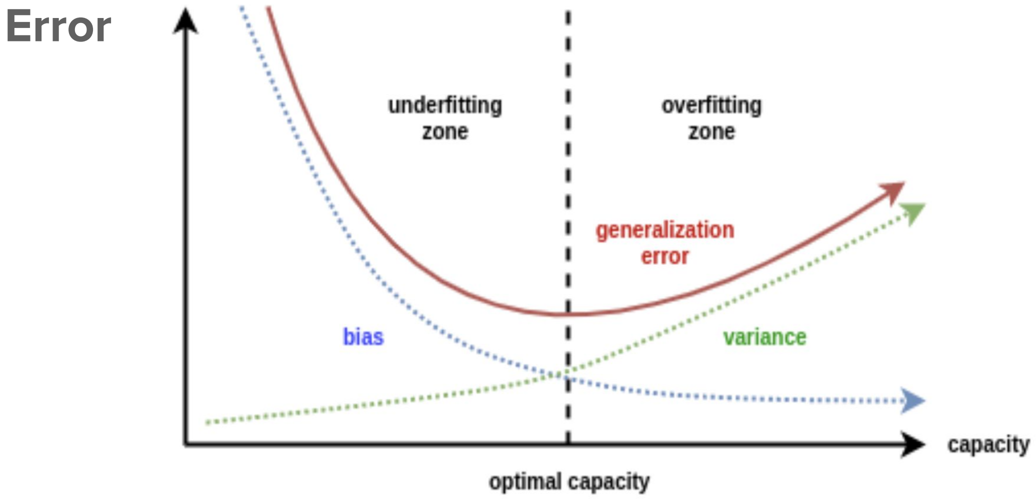
Variance

- Complex models can overfit
  - Fites the noise in the data





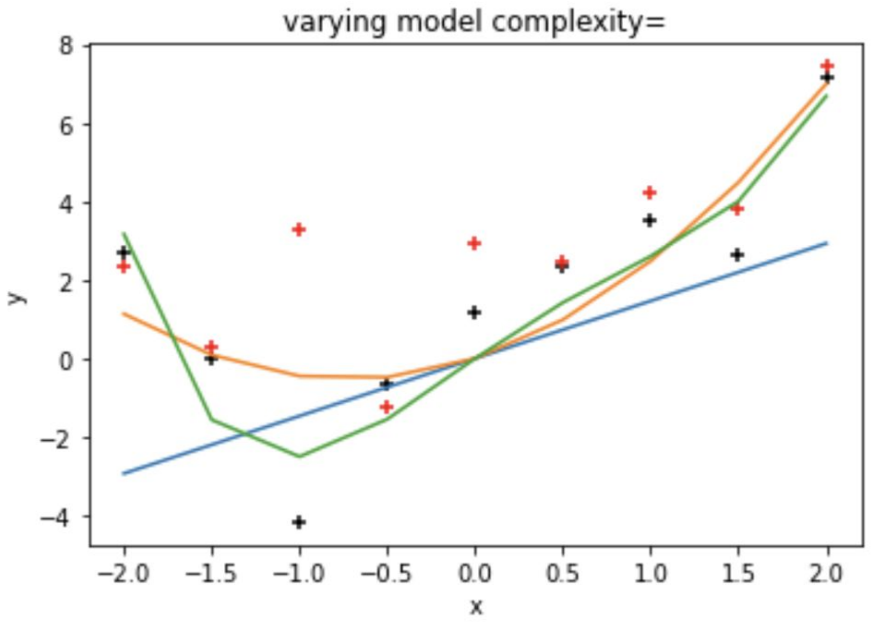
Picking the right model complexity



https://images.deepai.org/

The “right” complexity generalizes best

- Too simple a model “underfits”
  - It fails to capture the signal in the data
- Too complex a model “overfits”
  - It fits the noise in the data, aad so generali



liinear, quadratic and quartic models
train and test data

colab

Generalization and Overfitting

- Deep learning often uses more parameters than observations
  - ‘should’ massively overfit
- Deep learning on images (Zhang, Bengio, Hardt, Recht, and Vinyalsn 2017)
  - gives 0 training error -- and small test error
  - gives 0 training error with randomized labels
- GPT-3 (175B params trained on 500B words) seems to memorize a lot
  - Q. What do you call a droid that takes the long way around?
  - A. R2 detour.

Regularization is key

- **Modern neural nets don’t overfit as much as one might expect**
  - often trained with similar numbers of weights and observations
- **Best accuracy from giant networks with lots of regularization**
  - best: combine many different regularizations

Today: Regularization

- Regularization controls overfitting in overparameterized models
- Regularization by
  - shrinking: L1, L2, early stopping
  - data augmentation
  - SGD
  - dropout
- Hyperparameter tuning is critical and expensive
- Adversarial attacks
  - defense via regularization

Overparameterization and Overfitting

Lyle Ungar

neuromatch academy

Overparameterization

- If you have more adjustable parameters than you have observations, you can generally fit the training data perfectly
  - E.g. 100 patients, each with an image with 40,000 voxels
- What happens when to try to use linear regression with more features than observations?

Regularization is Shrinkage

Lyle Ungar

neuromatch academy

Shrinkage is Regularization

- We'll see many ways to shrink parameters
  - L2 penalties
  - set some of them to zero
- Smaller weights
  - *regularize* more
  - give smoother models
  - give models with lower "capacity"
  - fit the noise less well

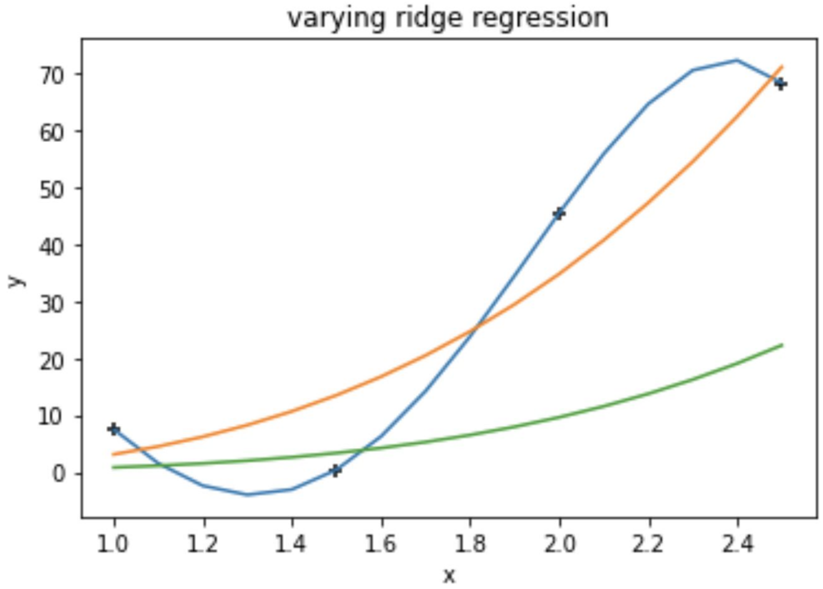
Shrinkage is Regularization

Fit $$y = c_0 + c_1 x + c_2 x^2 + c_3 x^3$$

with a ridge penalty of 0, 0.5, or 5,000

˙higher penalty = smoother prediction

colab



Shrinkage is Regularization

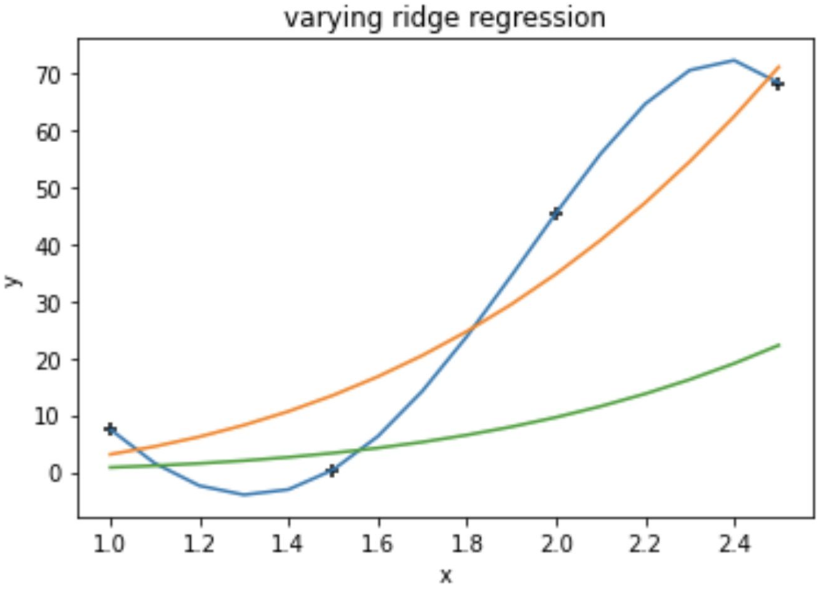
Fit $$y = c_0 + c_1 x + c_2 x^2 + c_3 x^3$$

with a ridge penalty of 0, 0.5, or 5,000

$||\mathbf{c}||_2$ 518, 17 0.5

higher penalty = smaller weights

colab



Generalization and overfitting

Lyle Ungar

neuromatch academy

Generalization and Overfitting

- Deep learning often uses more parameters than observations
    - ‘should’ massively overfit
- Deep learning on CIFAR10 and IMAGENET
    - gives 0 training error -- and small test error
    - gives 0 training error with randomized labels

Zhang, Bengio, Hardt, Recht, and Vinyalsn 2017

Regularization via early stopping

Lyle Ungar

neuromatch academy

Early stopping

- Initialize with small weights
- These get bigger as you do gradient descent
- Stop when they are the ‘optimal’ size