https://ttpoll.com/p/548136

COSC 325: Introduction to Machine Learning

Dr. Hector Santos-Villalobos

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Lecture 10: Bias, Variance, and Regularization

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

Class Announcements

Homework:

Homework #3 DD 09/29
Start early. TA support during weekends is not expected.

Course Project:

PRFAQ is due this Friday.
Check Additional Approved Datasets in the Course Project Assignment section.
Added two examples PRFAQ Examples.

Lectures:

On October 1st, no attendance record due to the Engineering Expo

Exams:

Exam #1: Thursday, 10/03

- Online
- Window 11 am to 1 pm
- 75 mins

Course grade distribution change:

- **Exams:** ~~45%~~ **35%**
- **Homework:** ~~20%~~ **30%**

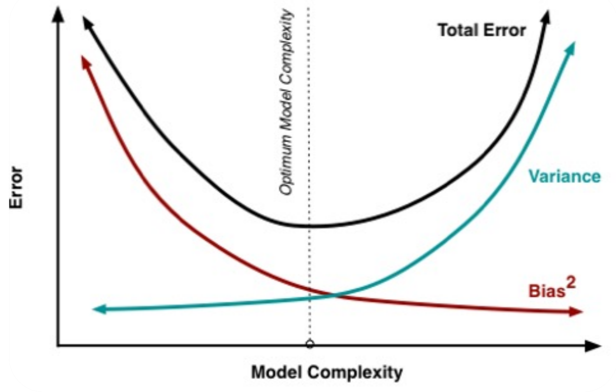
THE UNIVERSITY OF TENNESSEE KNOXVILLE

Review

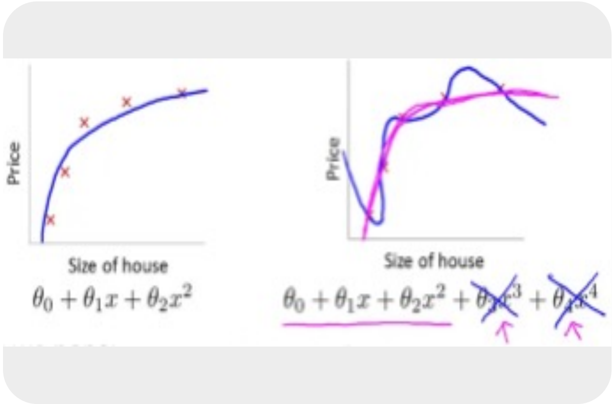
- Vectorized GD for logistic regression classification.
- Model evaluation
  - Dataset split
    - Training, validation, and testing
    - Random sampling while avoiding data leaks
  - Capacity
  - Overfitting

Today’s Topics

Variance and Bias



Regularization



THE UNIVERSITY OF TENNESSEE KNOXVILLE

Model Capacity

Capacity: the ability of a model to represent a wide variety of functions that map input data to output predictions. Also known as model complexity.

$$\mathcal{H} = \{h(X): X \to y\},$$

where $\mathcal{H}$ is the hypothesis space, which consists of all possible functions that the model $h(X)$ can learn on its architecture and parameters

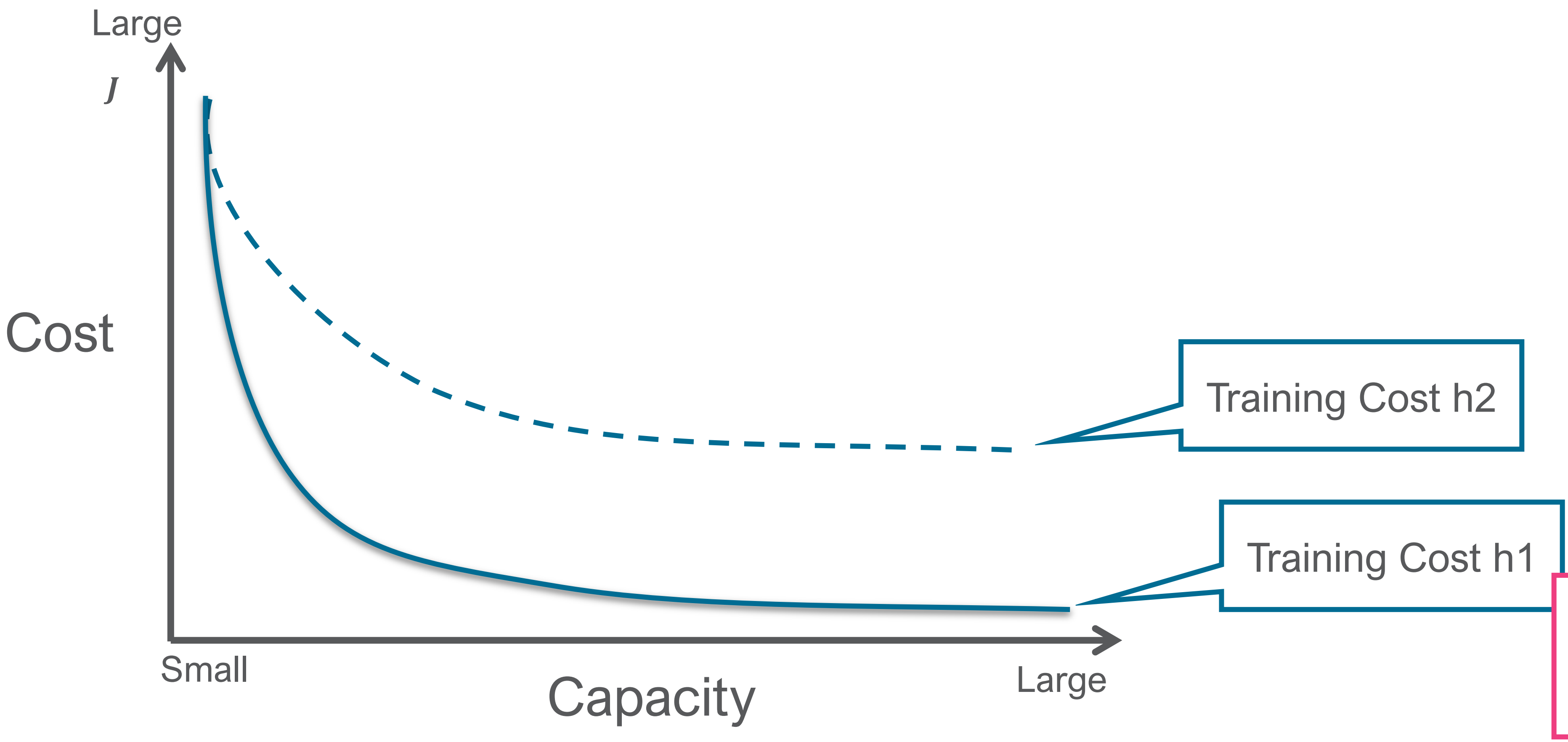
- Low capacity models (e.g., linear models) have smaller hypothesis space and can only represent simpler functions.
- High capacity models (e.g., deep neural networks) have a larger hypothesis space, enabling them to approximate more complex functions.

Overfitting and Underfitting

- ***Underfitting:*** both the training and validation errors are large.
  - Usually, the result of a low-capacity model
- ***Overfitting:*** gap between training and validation error
  - Validation error >> Training Error
- For a large hypothesis space being searched by a learning algorithm, there is a high tendency to ______fit

$$\mathcal{H} = \{h(X): X \rightarrow y\}, \text{ where } \mathcal{H} \text{ is very large}$$

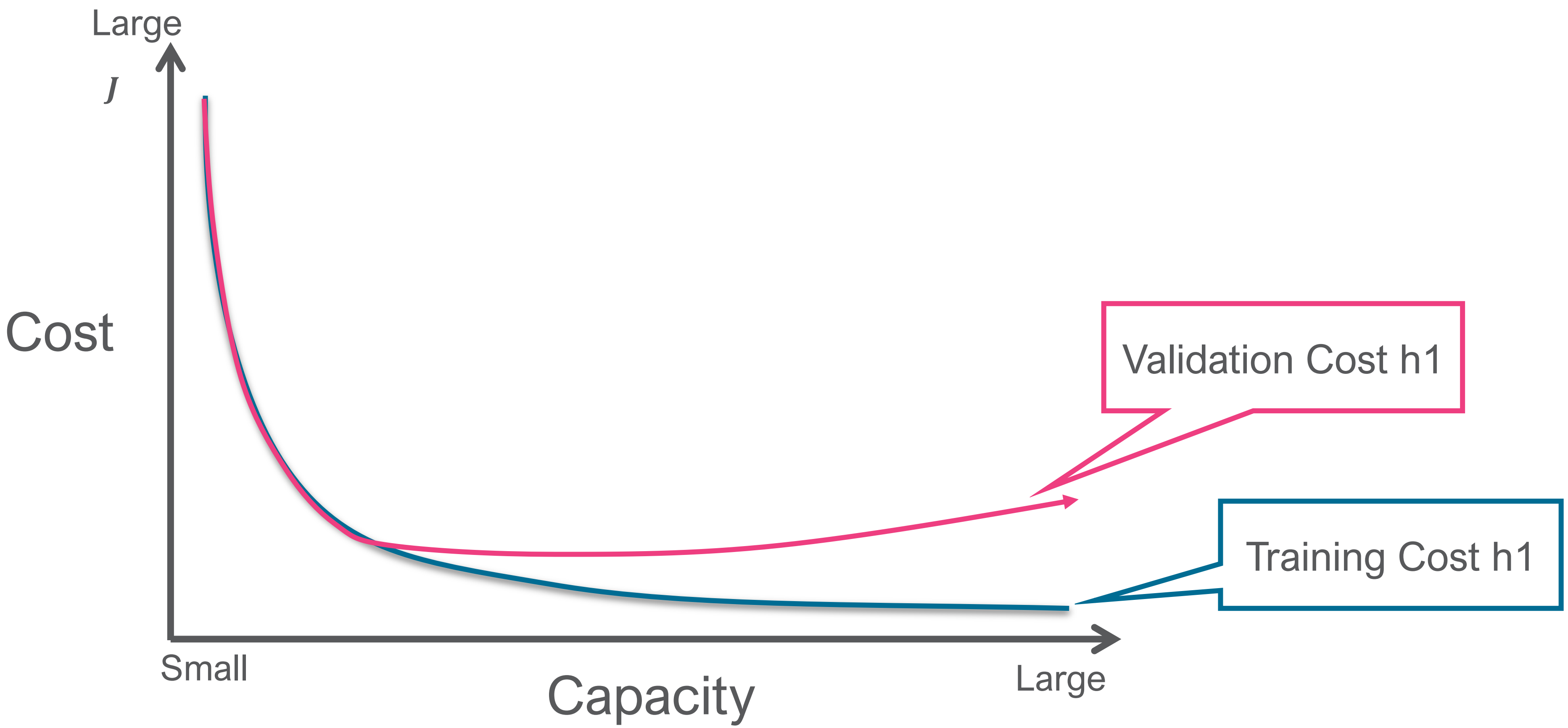
Overfitting and Underfitting



h_2 is _____fitting

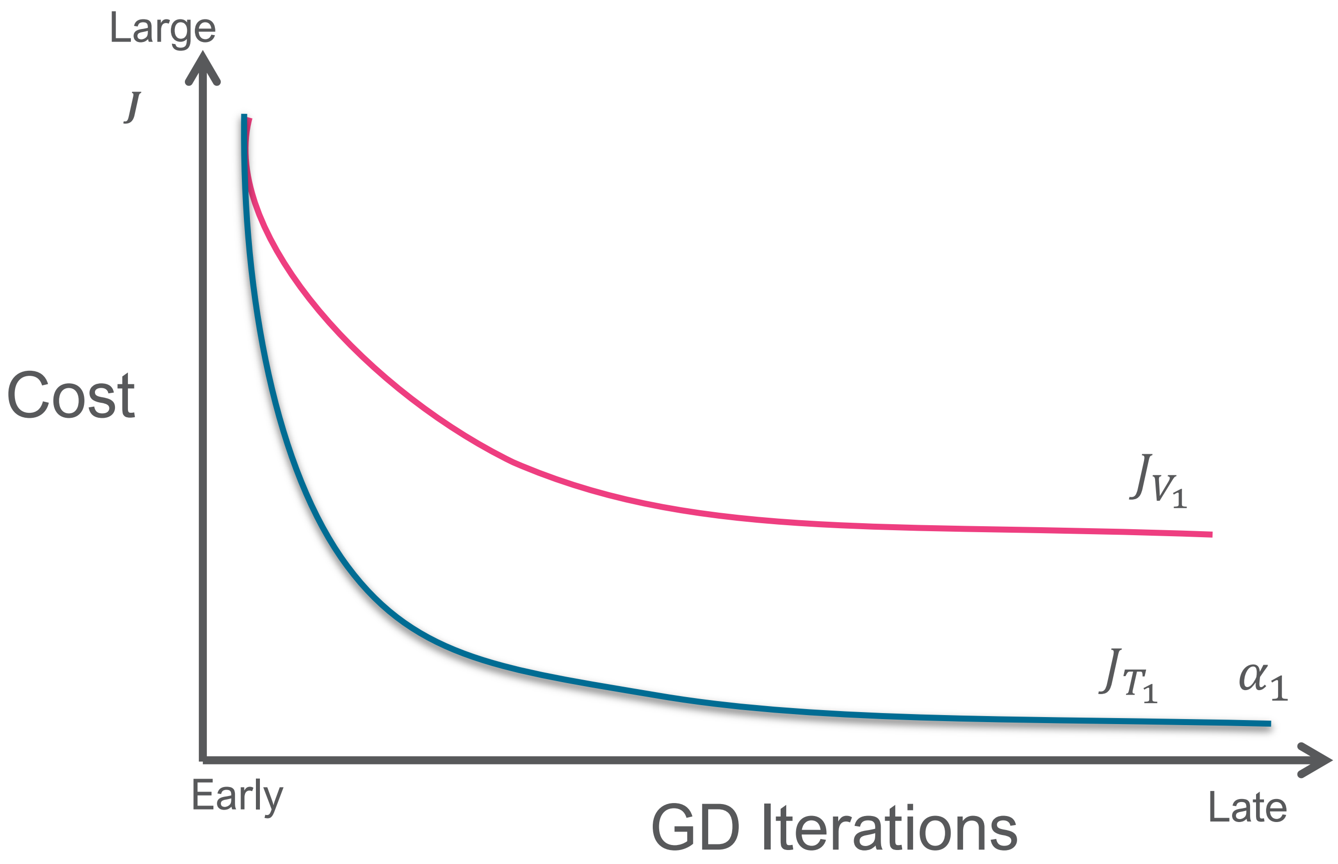
Can we say anything about h_1?

Overfitting and Underfitting



h_1 is overfitting

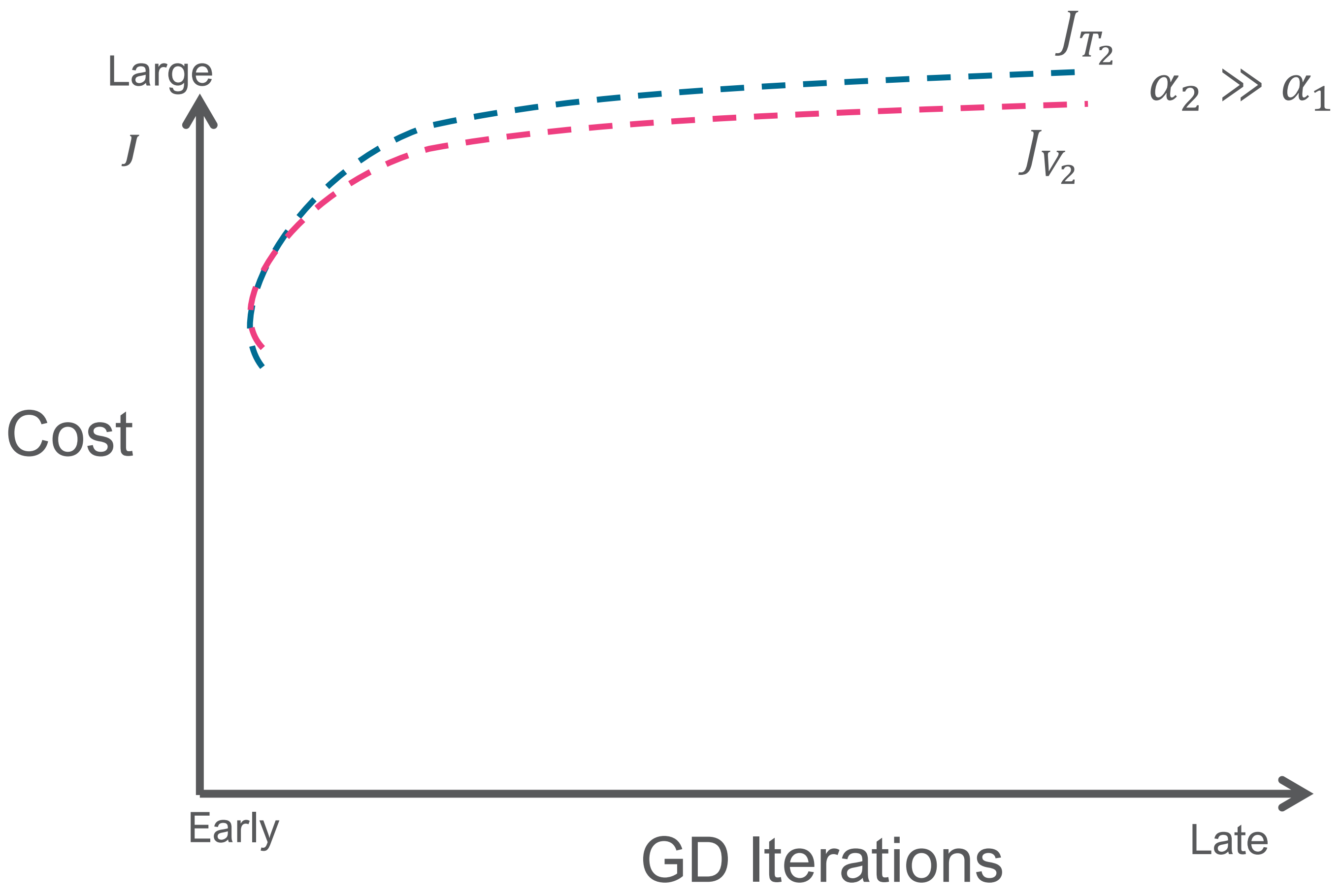
Relation to GD Iterations



Let's assume we want to pick a good learning rate α

T_1 is overfitting

Relation to GD Iterations

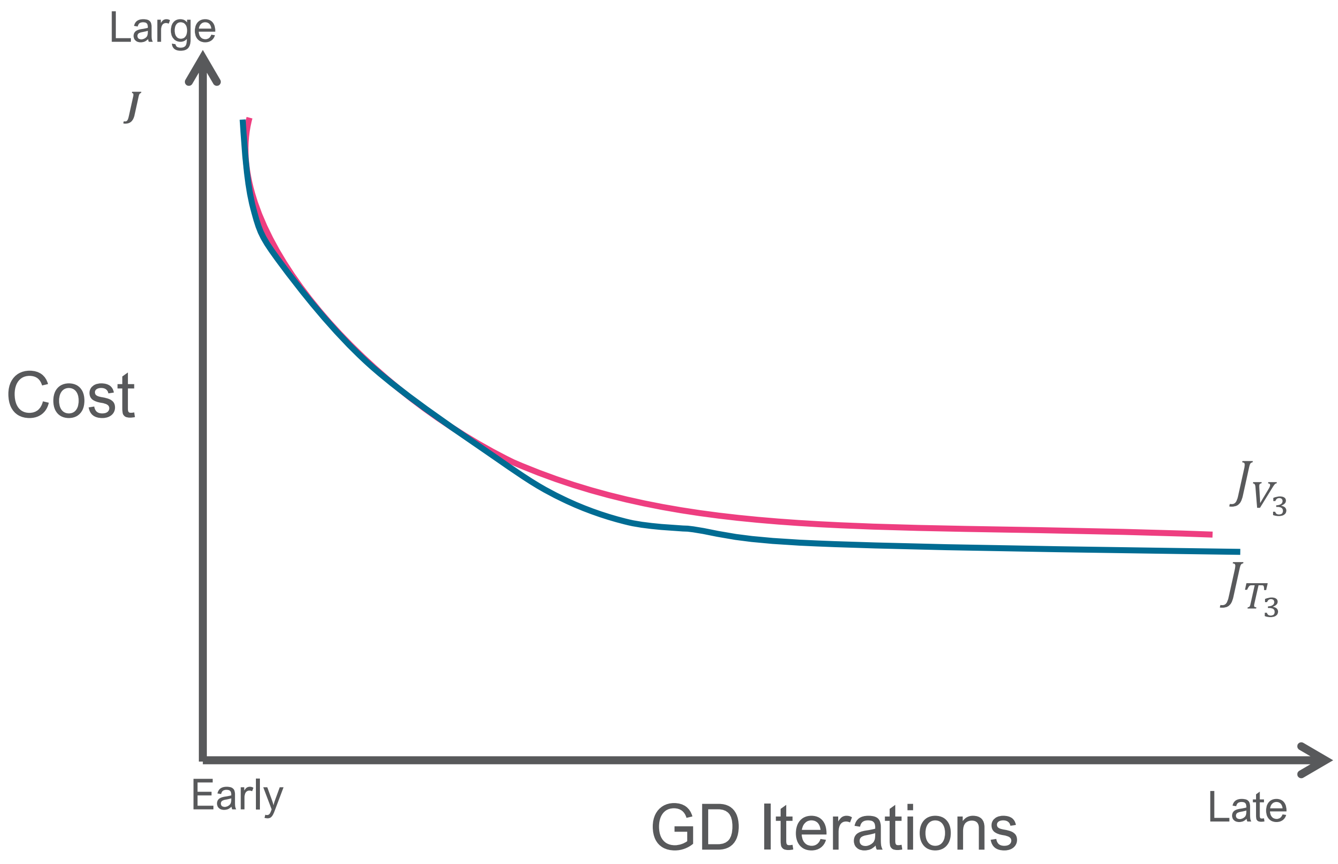


Let's assume we want to pick a good learning rate α

T_2 is underfitting

Relation to GD Iterations

Let's assume we want to pick a good learning rate α



$\alpha_2 > \alpha_3 > \alpha_1$

T_3 is about right

Model error/loss

$$E[ModelError] = Bias^2 + Variance + IrreducibleError$$

$$E[\mathcal{L}(y, \hat{y})]$$

Bias and Variance

- Sources of error
  - Bias
  - Variance
  - ***Irreducible Error***

Error that we cannot mitigate regardless of the architecture or algorithm.

Types

- Noise: imprecise measurements (mm vs km), faulty sensors, rounding error
  - E.g., miscalibration of X-ray CT scanner.
- Missing features
  - E.g., Predict house price without sqft of house
- Nondeterministic systems
  - E.g., human behavior, weather prediction
- Data labeling errors
  - E.g., subjective interpretation of labeling task

Bias and Variance

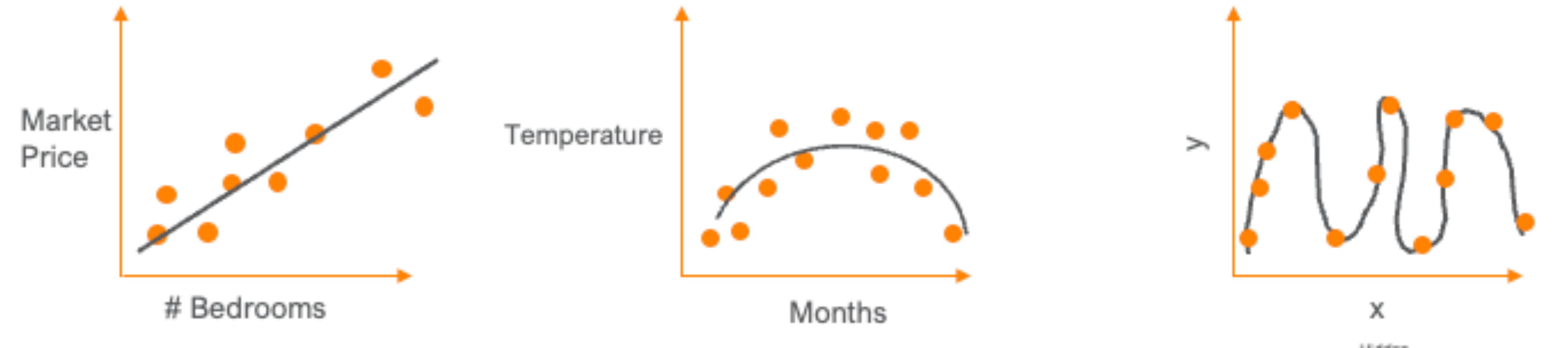
- Sources of error
  - ***Bias***
  - Variance
  - Irreducible Error

The deviation of the predictions from the ground truth.

Flexibility of the model to fit the true function perfectly.

$$Bias \propto \frac{1}{Flexibility}$$

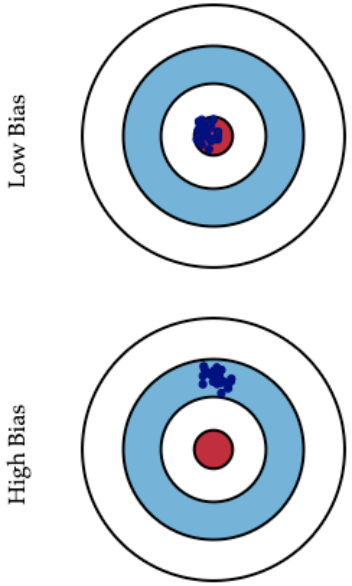
High Bias



Low Bias

Bias and Variance

- Sources of error
  - ***Bias***
  - Variance
  - Irreducible Error

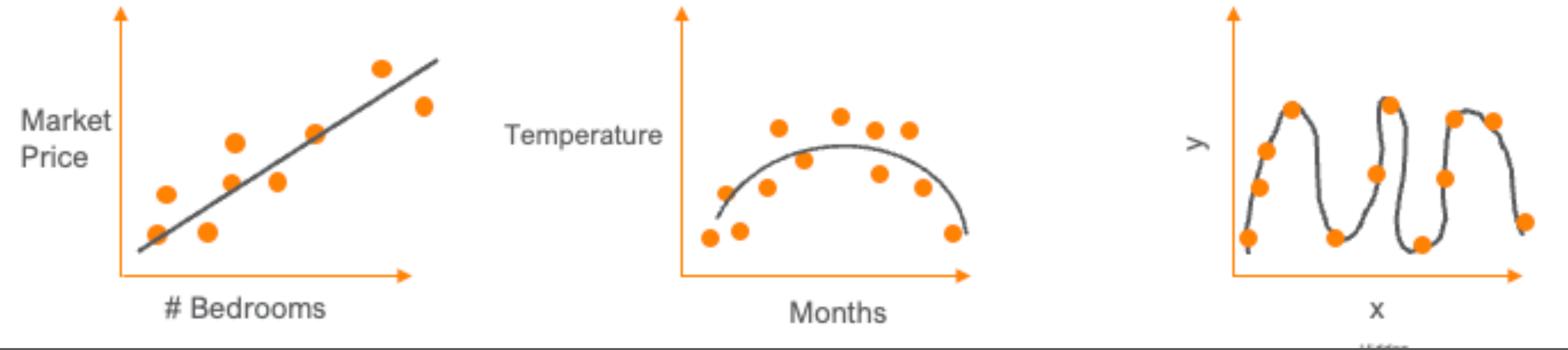


$$Bias^2 = \mathcal{L}(y, E[\hat{y}])$$

Bias^2 for SSE Loss

$$\mathcal{L}_{SSE}(y, E[\hat{y}]) = (y - E[\hat{y}])^2$$

High Bias



Low Bias

Bias and Variance

- Sources of error
  - Bias
  - ***Variance***
  - Irreducible Error

How much do my predictions change when I change my data (e.g., training vs validation sets)?

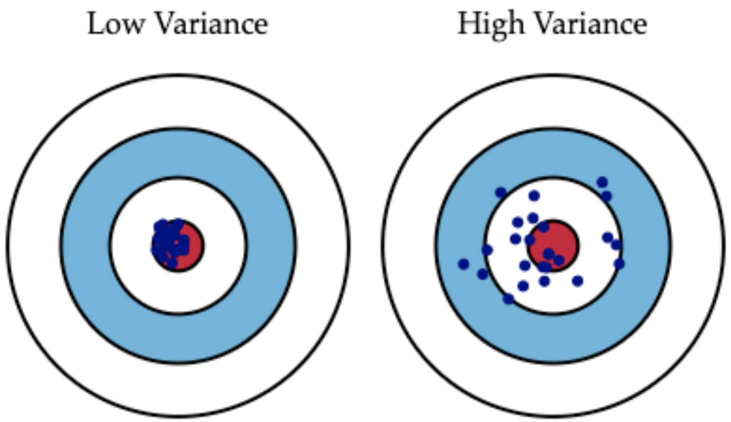
The deviation between models' predictions.

Usually, $bias \propto \frac{1}{variance}$

Exceptions with the latest DL architectures.

Bias and Variance

- Sources of error
  - Bias
  - ***Variance***
  - Irreducible Error



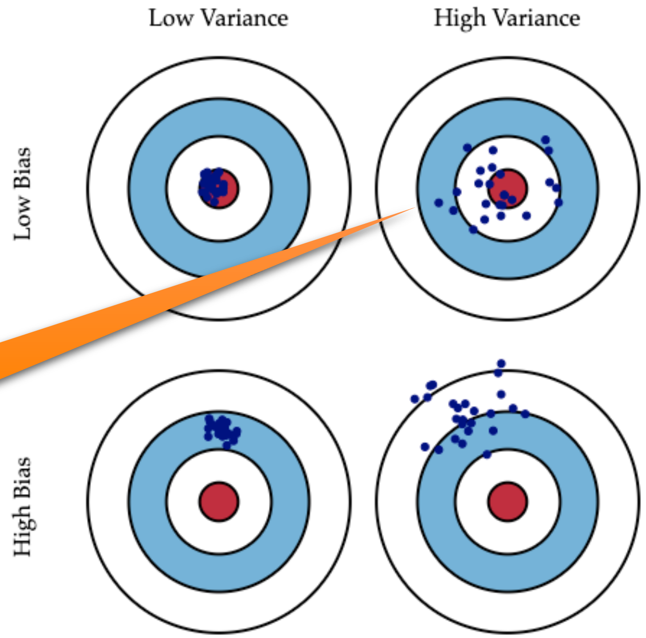
$$Variance = E[\mathcal{L}(\hat{y}, E[\hat{y}])]$$

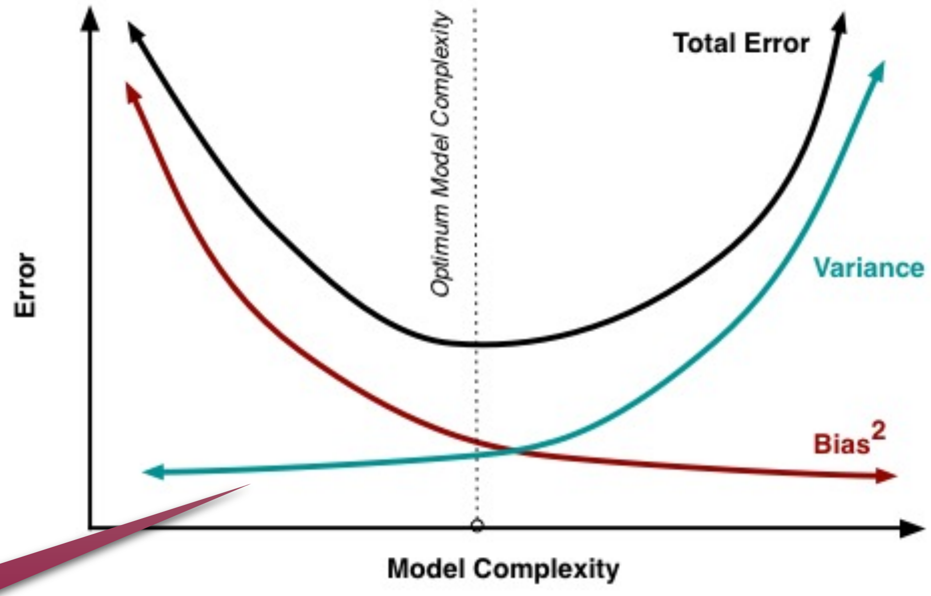
Variance for SSE Loss

$$\mathrm{E}[\mathcal{L}_{SSE}(\hat{y}, E[\hat{y}])] = E[(E[\hat{y}] - \hat{y})^2]$$

Bias and Variance

- Sources of error
  - Bias
  - Variance
  - Irreducible Error





Should we optimize for low bias?

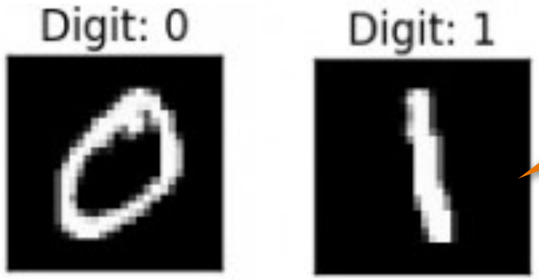
No, both are equally important.

Source figures: https://scott.fortmann-roe.com/docs/BiasVariance.html

Inspecting Bias and Variance from Cost

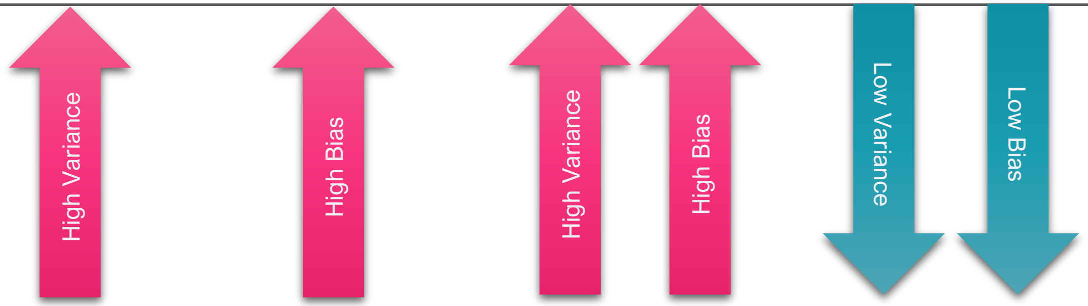
- Datasets
  - Training
    - Train model parameters
  - Validation
    - Test model parameters
  - Test
    - Test model parameters and ***final*** hyperparameters

Task:



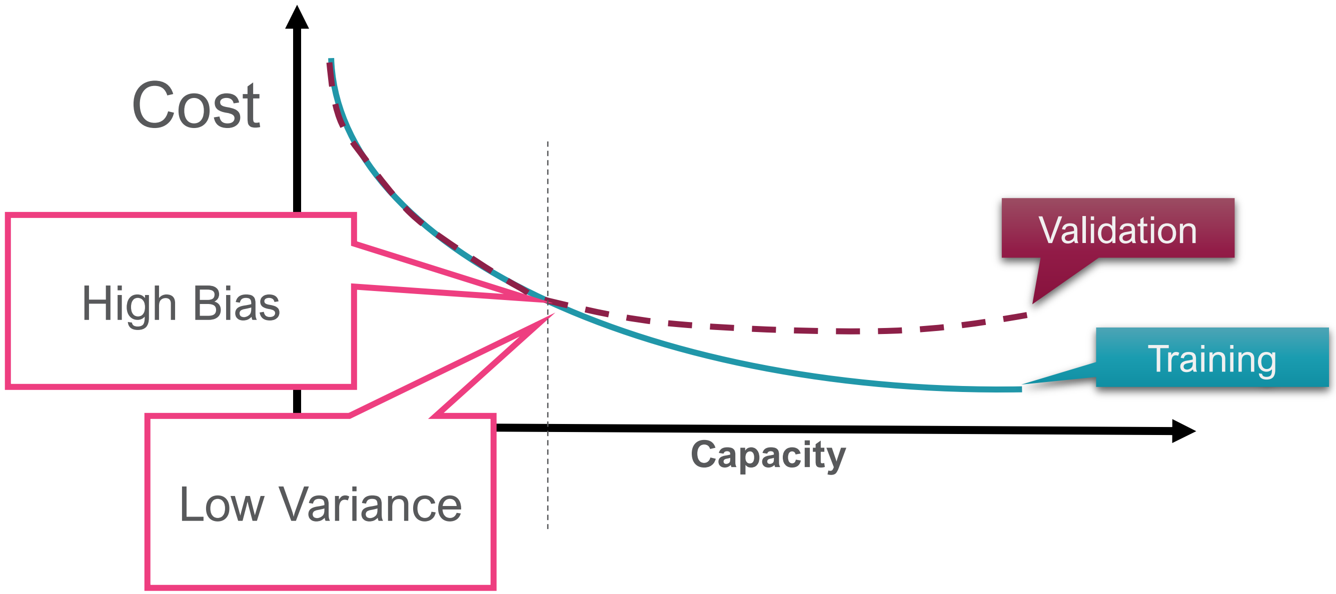
This task is easy for humans—less than 1% error.

	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$
Training Error	0.01	0.10	0.12	0.005
Validation Error	0.10	0.11	0.25	0.01



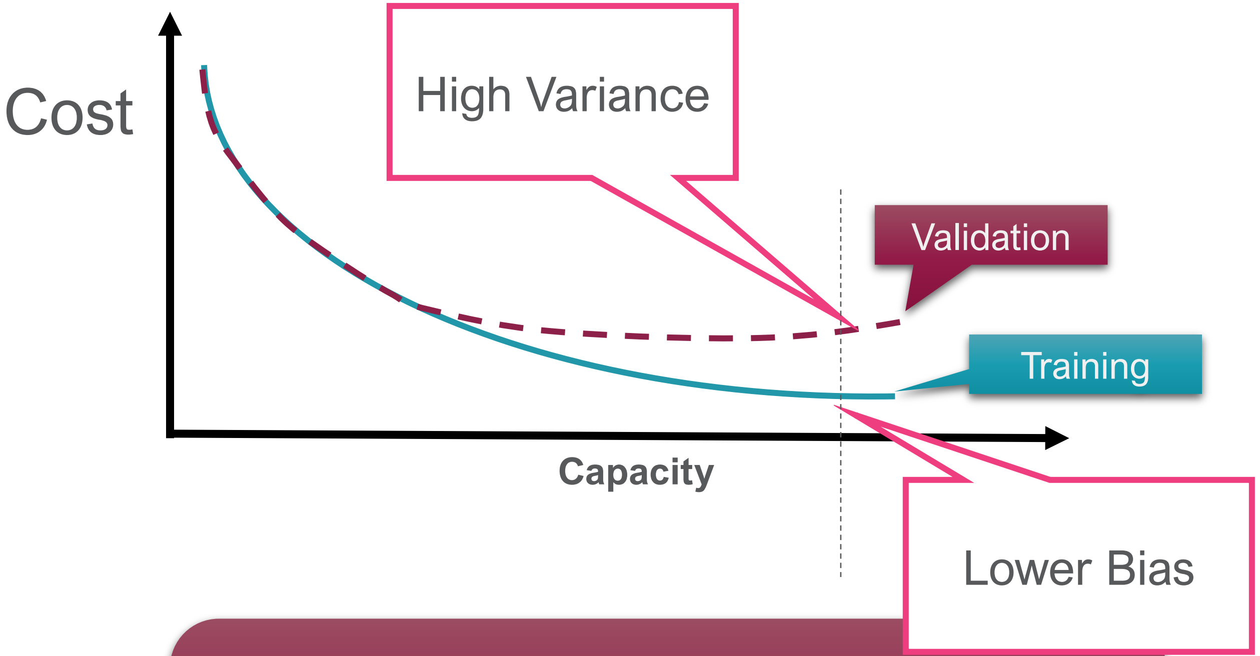
Underfitting

- Datasets
  - Training
    - Train model parameters
  - Validation
    - Test model parameters
  - Test
    - Test model parameters and ***final*** hyperparameters



Overfitting

- Datasets
  - Training
    - Train model parameters
  - Validation
    - Test model parameters
  - Test
    - Test model parameters and ***final*** hyperparameters



Regularization techniques address overfitting.

Notebook Time

So far

Dataset

Collect

Analyze

Process

Split

Train/Validate

Test

High Variance? Validation set.

- More data
- Regularization
- New mode hypothesis

High Bias? Training set.

- Bigger or more complex model
- Train for longer
- New model hypothesis

Pop Quiz

MULTIPLE CHOICE

A model with low bias and high variance tends to _______.

A. Overfit

B. Underfit

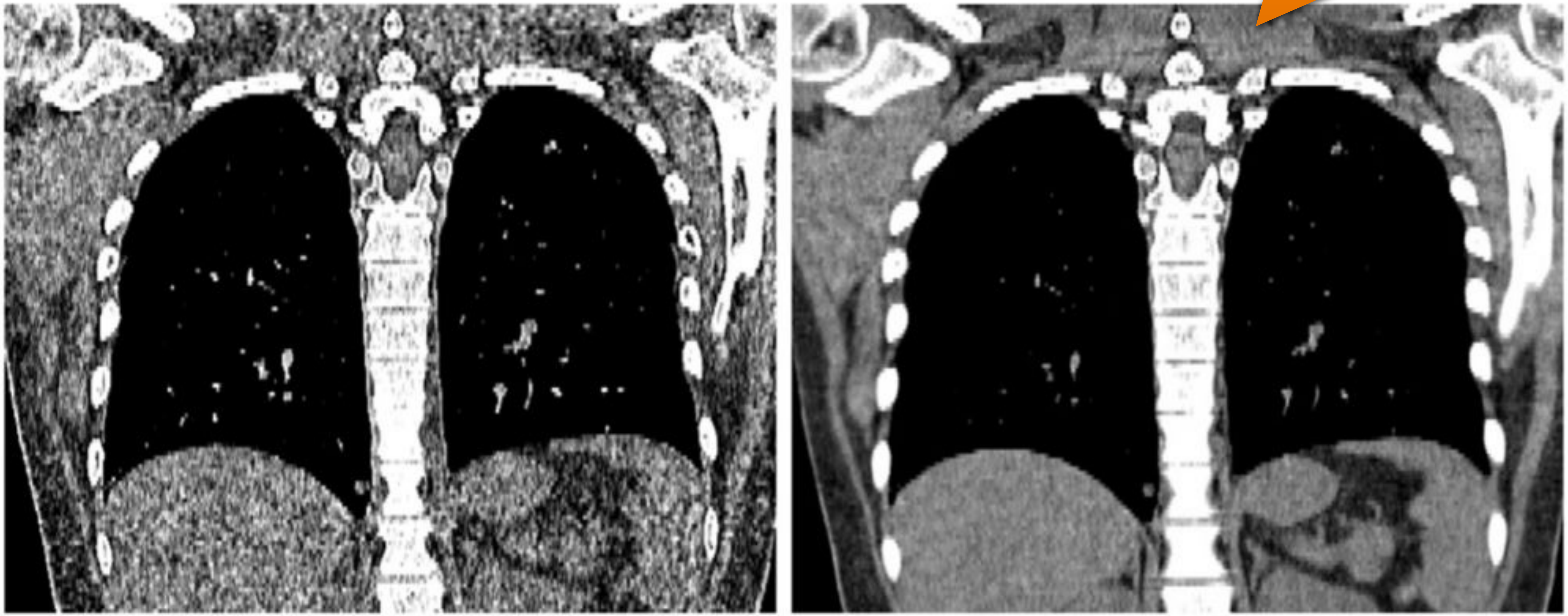
Regularization

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

What is regularization?

Q-Generalized Gaussian Markov Random Field (QGGMRF) Regularizer: Enhance details on low-contrast regions.

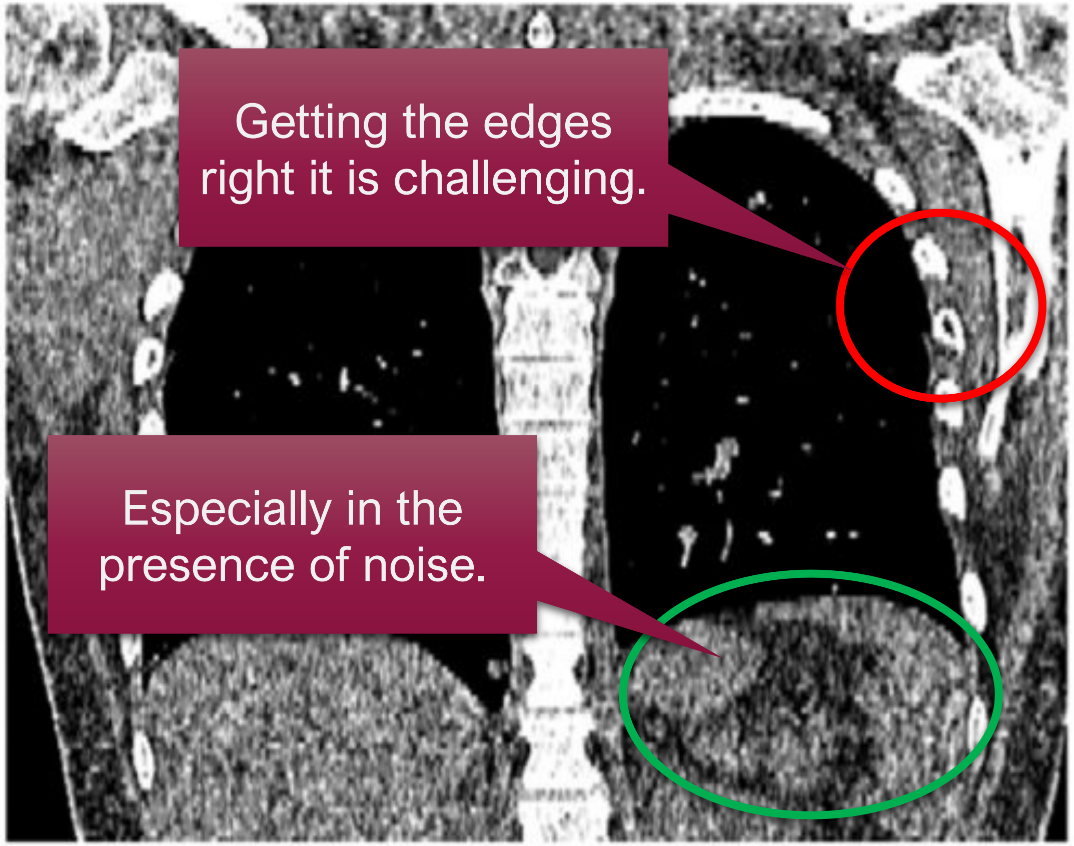
Medical CT coronal image



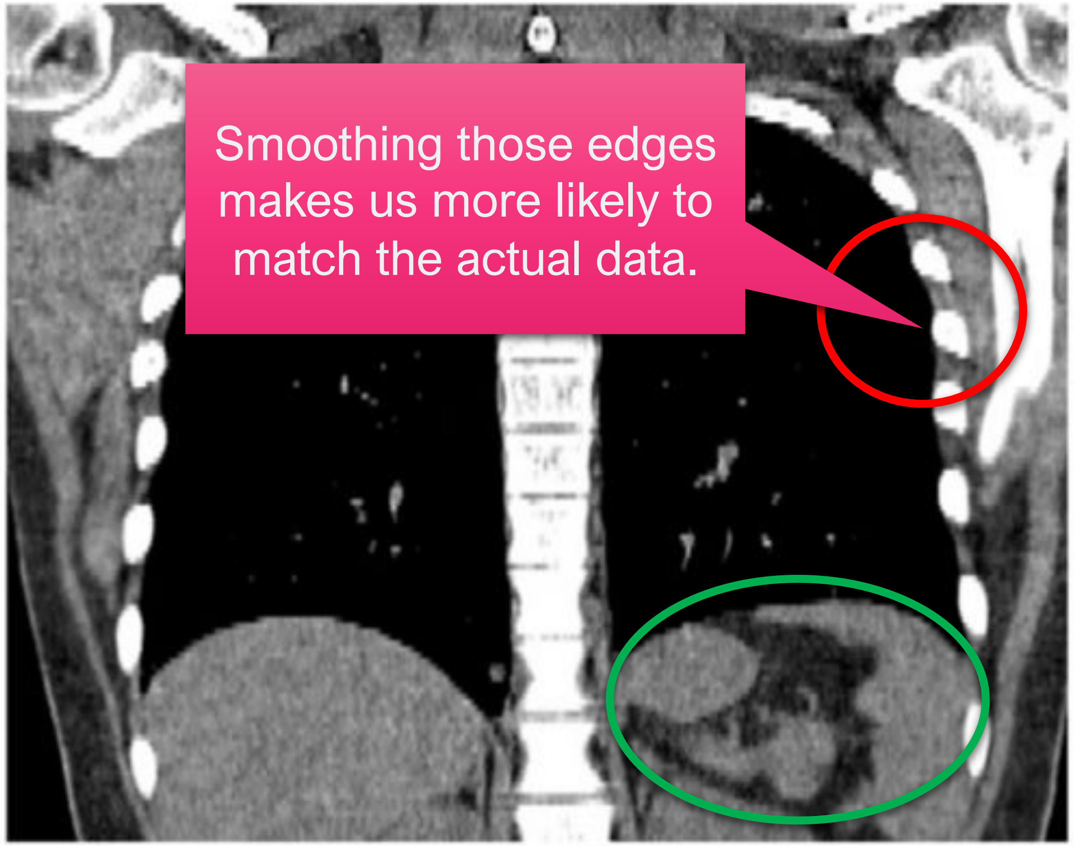
Siemens SOTA Reconstruction　　Model-Based Image Reconstruction (MBIR)

What is regularization?

Medical CT coronal image

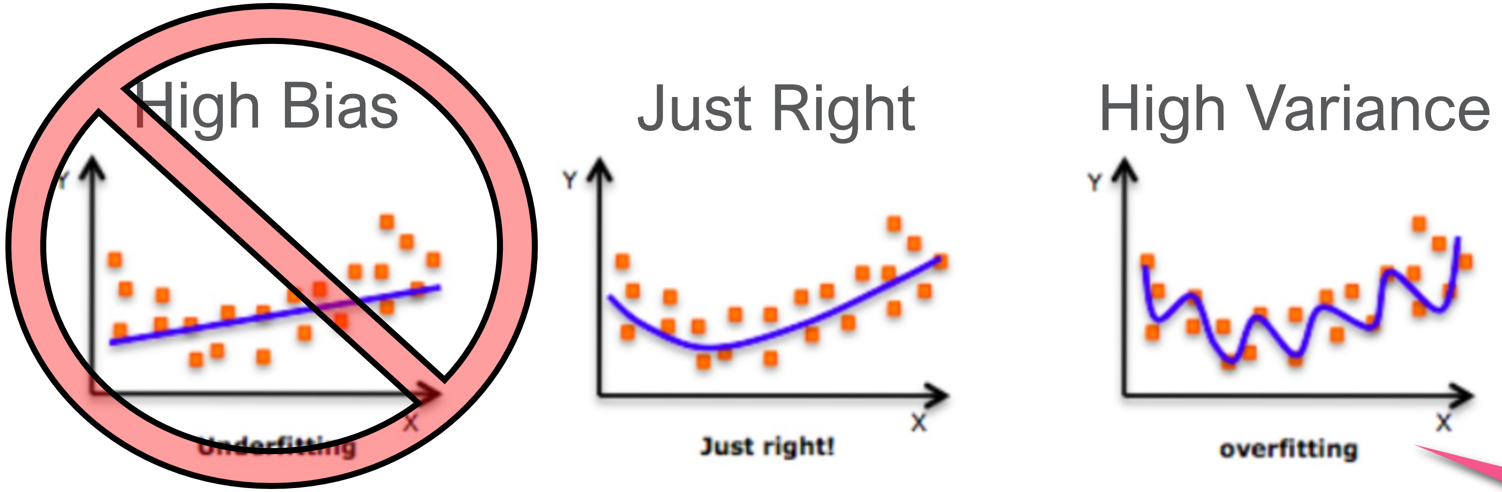


Siemens SOTA Reconstruction



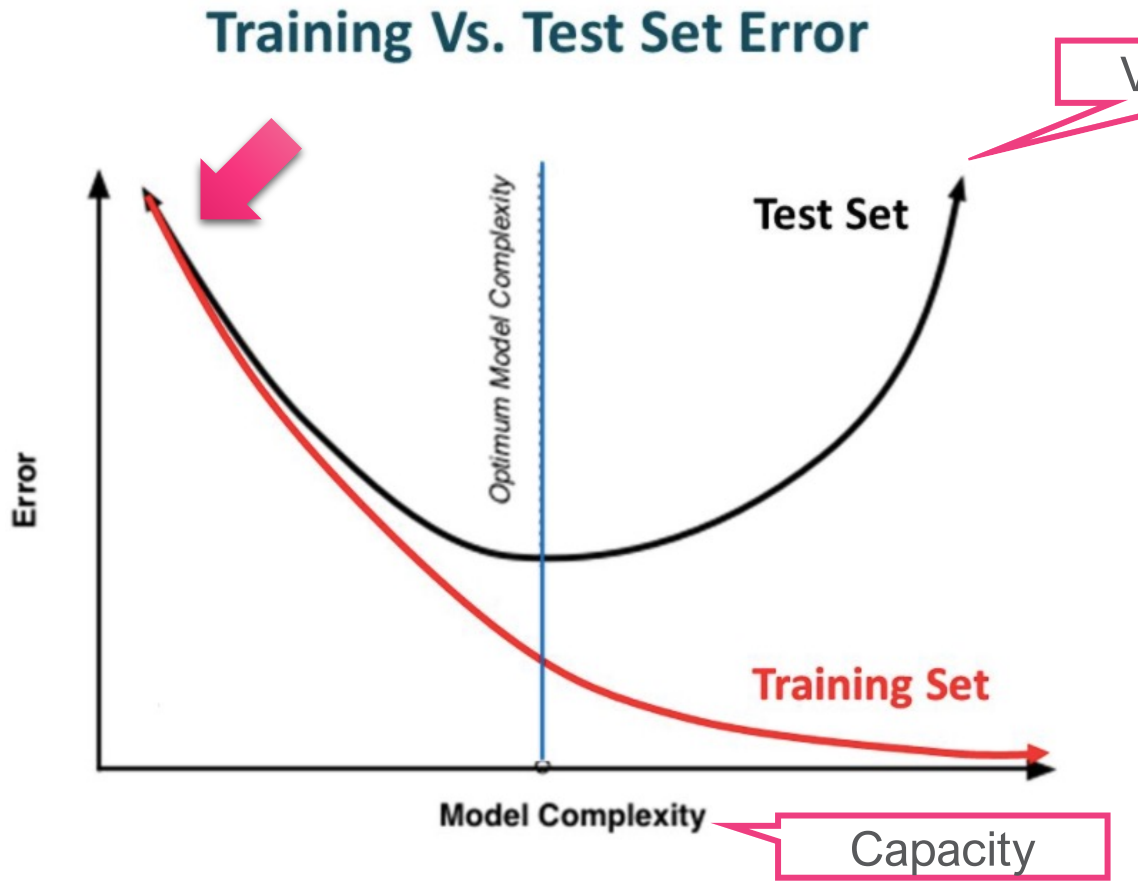
Model-Based Image Reconstruction (MBIR)

Regularization helps us find the "Just Right" spot.



Need help!

Regularization and Model Error



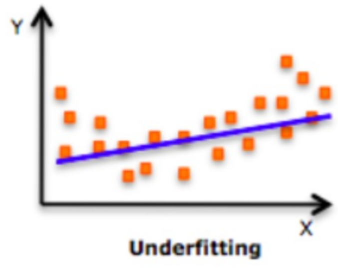
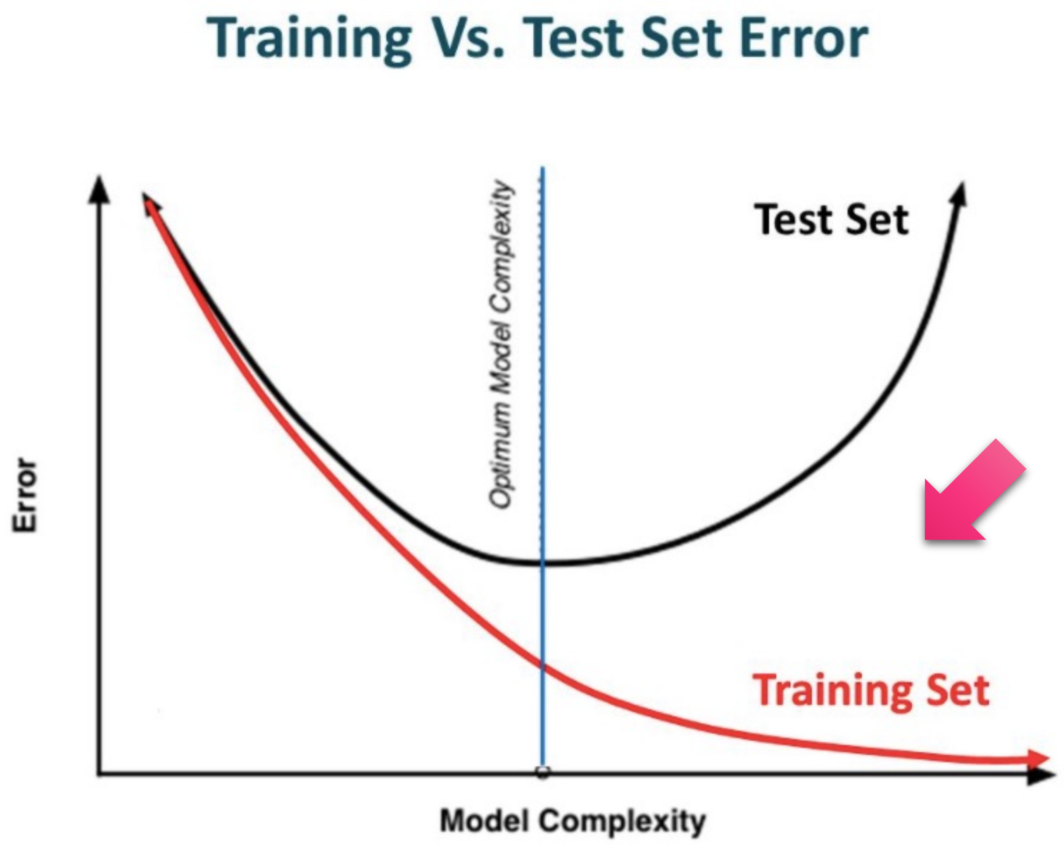
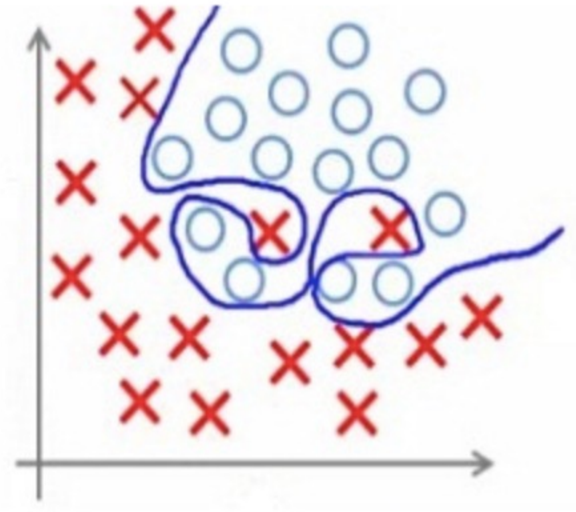
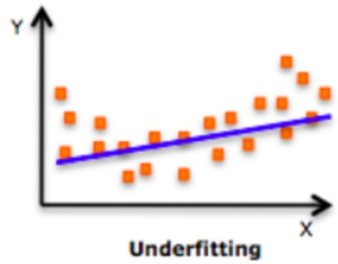


Image source: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

Regularization and Model Error







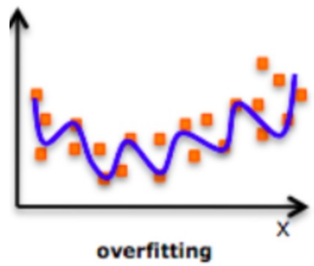
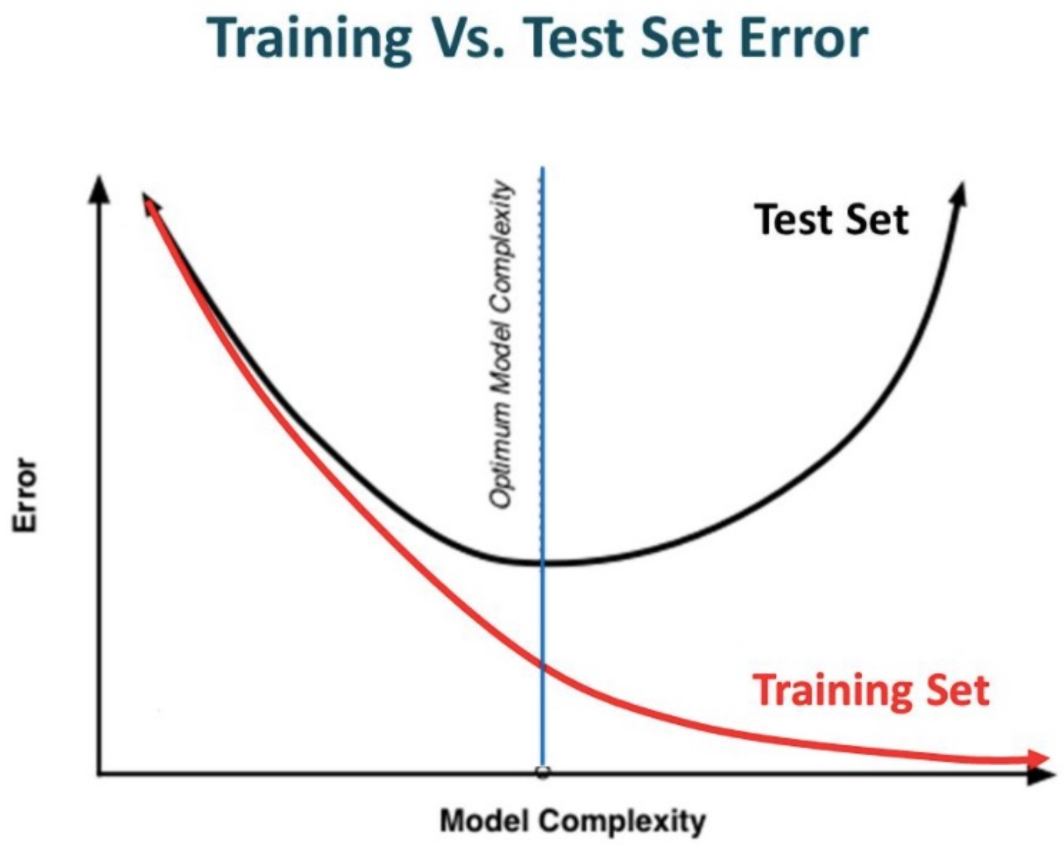
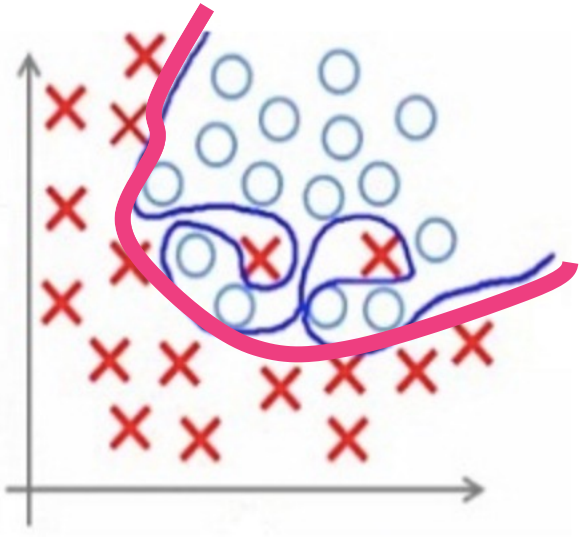


Image source: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

Regularization and Model Error



Regularization



Good-fitting

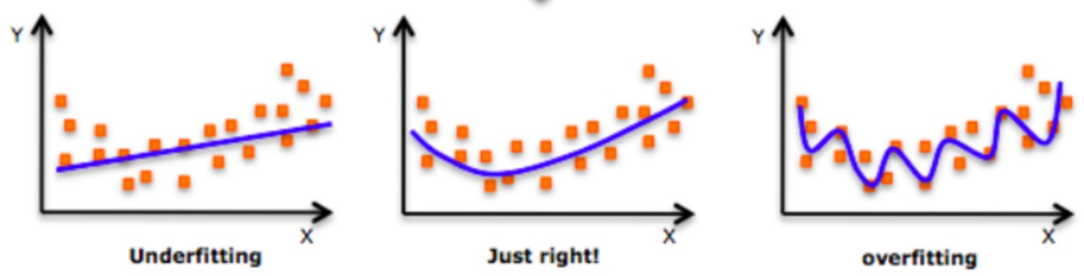


Image source: https://www.analyticsvidhya.com/blog/2018/04/fundamentals-deep-learning-regularization-techniques/

https://ttpoll.com/p/548136

L2 Regularization (Ridge)

Regularization

Parameter to control how much to regularize.

Removes sqrt of L2 norm.

Logistic Regression

$$\min_{w,b} J(w) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \underbrace{\frac{\lambda}{2m}\|w\|_2^2}$$

L2 Regularization: $\|w\|_2^2 = \sum_{j=1}^{m} w_j^2 = w^T w$

Popular approach.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

L2 Regularization (Ridge)

Logistic Regression

$$\min_{w,b} J(w) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) \quad + \frac{\lambda}{2m}\|w\|_2^2$$

L2 Regularization: $\|w\|_2^2 = \sum_{j=1}^{m} w_j^2 = w^T w$

Heavily penalizes larger weights

L2 Regularization (Ridge)

Logistic Regression

$$\min_{w,b} J(w) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) \quad + \frac{\lambda}{2m}\|w\|_2^2$$

L2 Regularization: $\|w\|_2^2 = \sum_{j=1}^{m} w_j^2 = w^T w$

Note: $\frac{\partial\left(\frac{\lambda}{2m}\|w\|_2^2\right)}{\partial w_i} = \frac{\partial\left(\frac{\lambda}{2m}\sum_{j=1}^{m} w_j^2\right)}{\partial w_i} = \frac{\lambda}{m} w_i$

Gradient Descent with Regularization

$$dW = \frac{dJ}{dW} = dW = \frac{1}{n} A \quad dZ + \frac{\lambda}{m} W$$

We are penalizing weight magnitude.

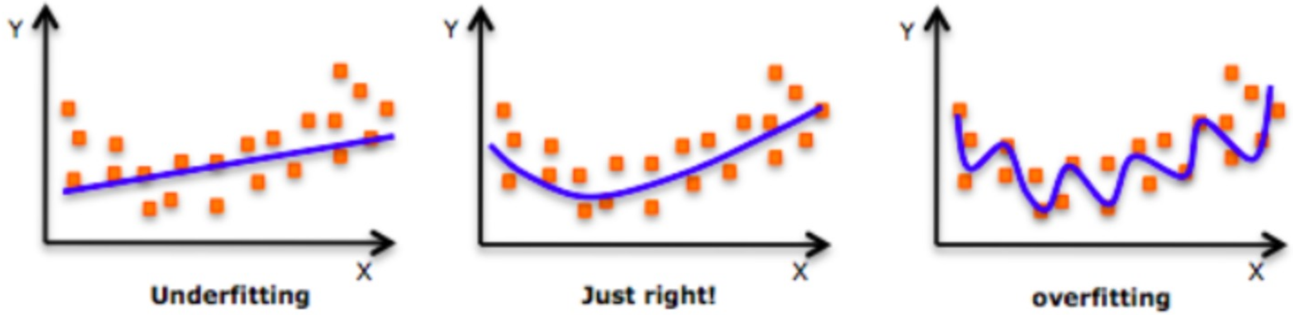
$$W := W - \alpha dW \qquad = W - \alpha \left[\frac{1}{n} A \quad dZ + \frac{\lambda}{m} W \right]$$

$$= \left(1 - \frac{\alpha\lambda}{m}\right) W - \alpha \left[\frac{1}{n} A \quad dZ \right]$$

Also called "Weight Decay" for this reason.

Intuition on Regularization

High Bias Just Right High Variance



$$\min_{W} J(W) = -\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\|W\|_2^2$$

Intuition on Regularization

$$\min_{W} J(W) = -\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\|W\|_2^2$$

If $\lambda \gg 0,\quad$ then $W \rightarrow 0$

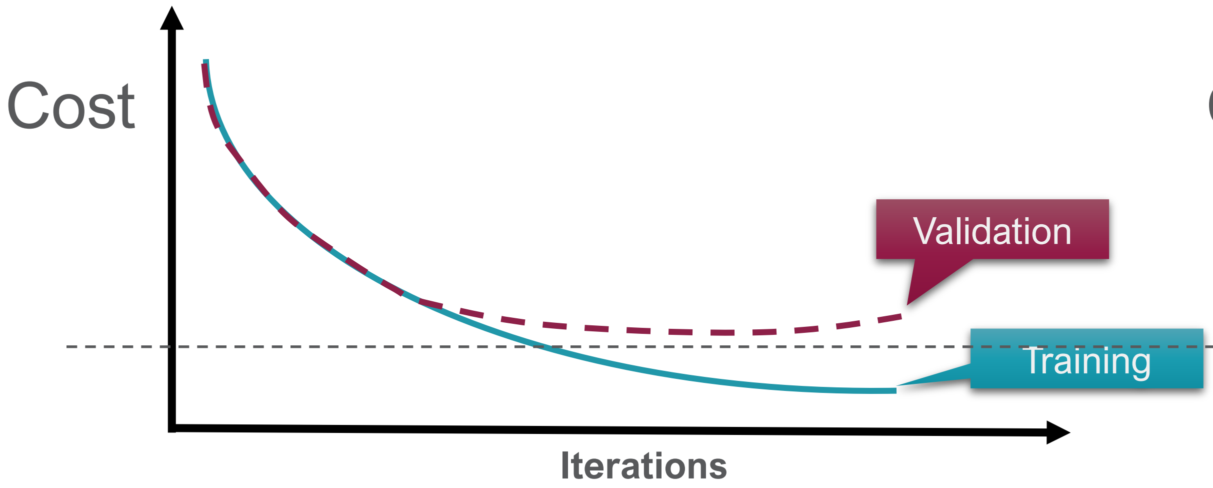
Intuition on Regularization

$$\min_W J(W) = -\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\|W\|_2^2$$

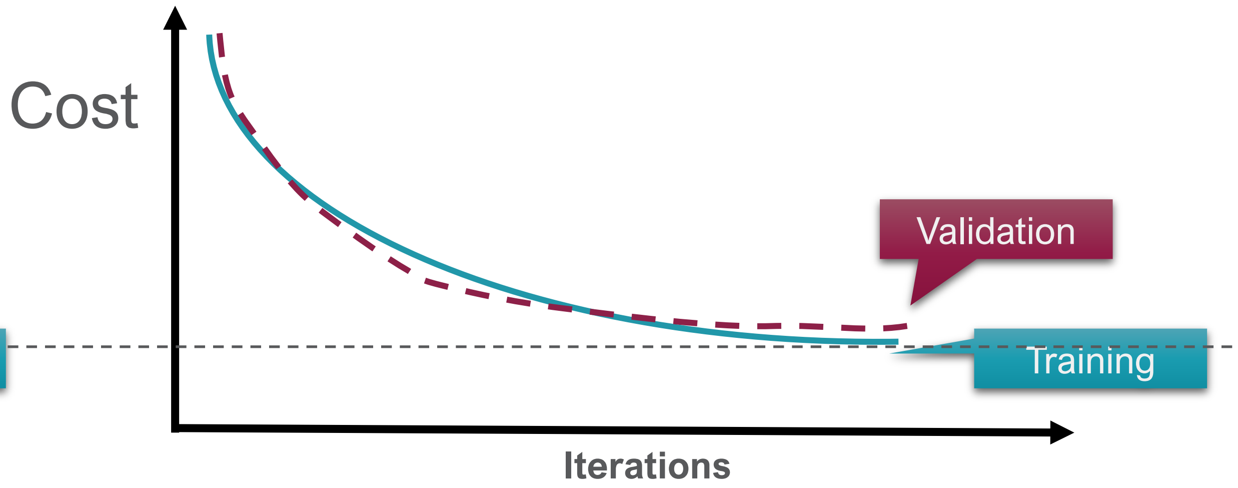
If $\lambda \gg 0$, then $W \to 0$

Overfitting

Low Bias-High Variance



Higher Bias-Low Variance



$$\min_{W} J(W) = -\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\left(\hat{y}^{(i)}, y^{(i)}\right) + \frac{\lambda}{2m}\|W\|_2^2$$

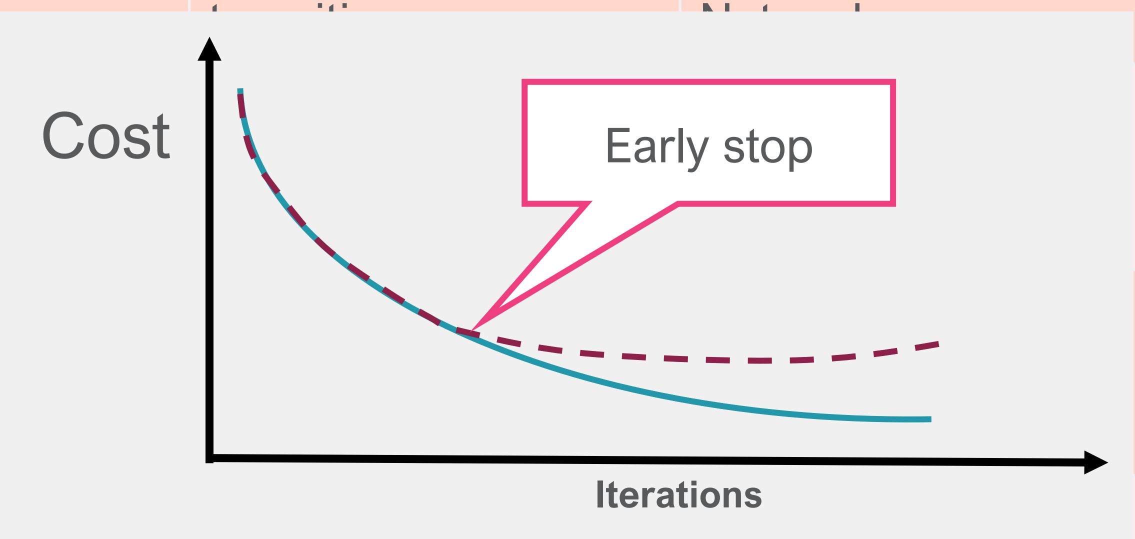
If $\lambda \gg 0$, then $W \rightarrow 0$

Popular Regularization/Penalty Terms

Technique	Formula	Type	Effect	Common use cases
Ridge (L2)	$\frac{\lambda}{2}\sum_{i=1}^{m} w_j^2 = w^T w$	Penalizes squared weights	Rewards smaller weights, smoother transitions.	Linear/Logistic Regression, Neural Networks
Lasso (L1)	$\lambda\sum_{j=1}^{m} \lvert w_j \rvert$	Penalizes absolute weights	Rewards sparsity (feature space reduction)	High-dimensional data
ElasticNet	$\frac{\lambda_1}{2}\lVert w\rVert_2^2 + \lambda_2\lVert w\rVert_1$	Combines Ridge and Lasso	Balances sparsity (L1) and smoothness (L2)	High-dimensional data with correlated features
Early Stopping	N/A	Stops training after specified cost event.	Prevents overfitting by using an earlier checkpoint.	Neural Networks

Popular Regularization/Penalty Terms

Technique	Formula	Type	Effect	Common use cases
Ridge (L2)	$\frac{\lambda}{2}\sum_{i=1}^{m} w_j^2 = w^T w$	Penalizes squared weights	Rewards smaller weights, smoother	Linear/Logistic Regression, Neural
Lasso (L1)	$\lambda\sum_{j=1}^{m}\lvert w_j\rvert$	Penalizes abs weights		
ElasticNet	$\frac{\lambda_1}{2}\lVert w\rVert_2^2 + \lambda_2\lVert w\rVert_1$	Combines Ri Lasso		
Early Stopping	N/A	Stops training after specified cost event.	by using an earlier checkpoint.	Neural Networks



Pop Quiz

The purpose of regularization is to ______________.

A. compute the average parameter/weight value.

B. decrease the likelihood of overfitting.

C. decrease the capacity of the model.

D. filter the outliers in the dataset.

Review

- Capacity
- Overfitting/Underfitting
- Bias-Variance Tradeoff
- Loss = Bias^2 + Variance + Irreducible Error
- Regularization techniques

Next Lecture

- Decision Trees
- Ensemble techniques