

# Machine Learning Basics

Changyou Chen

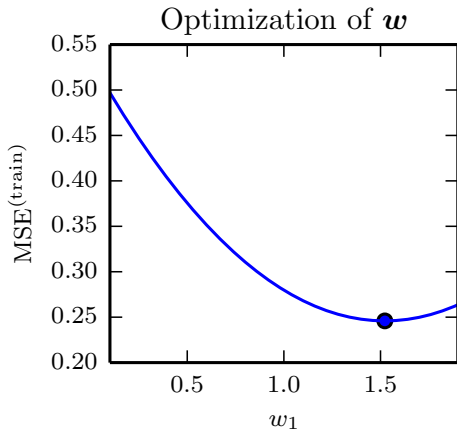
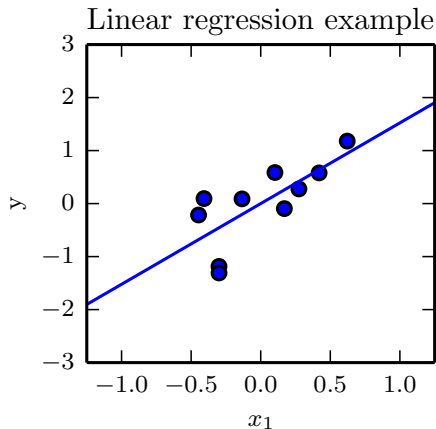
Department of Computer Science and Engineering  
University at Buffalo, SUNY  
changyou@buffalo.edu

February 7, 2019

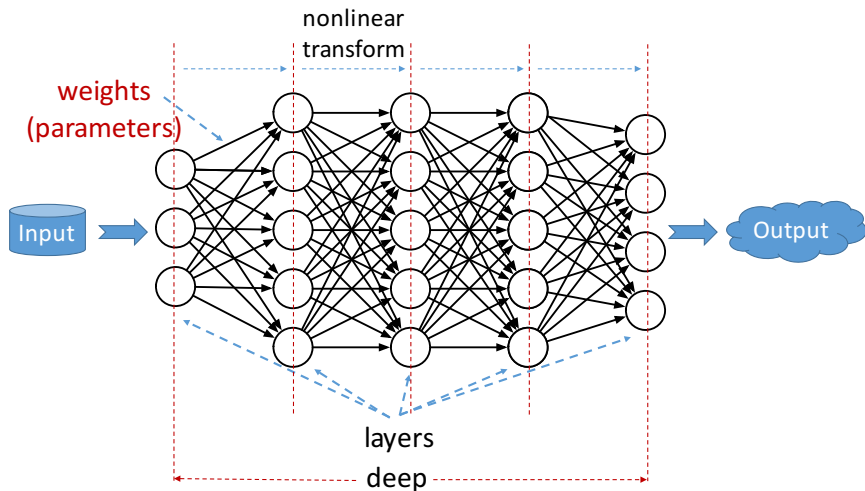
# Linear Regression

Find a linear model to fit the observed data  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$

Objective:  $\mathbf{w} = \arg \min_{\mathbf{w}} \sum_{i=1}^N |\mathbf{w}^T \mathbf{x}_i - \mathbf{y}_i|^2$



# Deep Neural Networks are Nonlinear Models

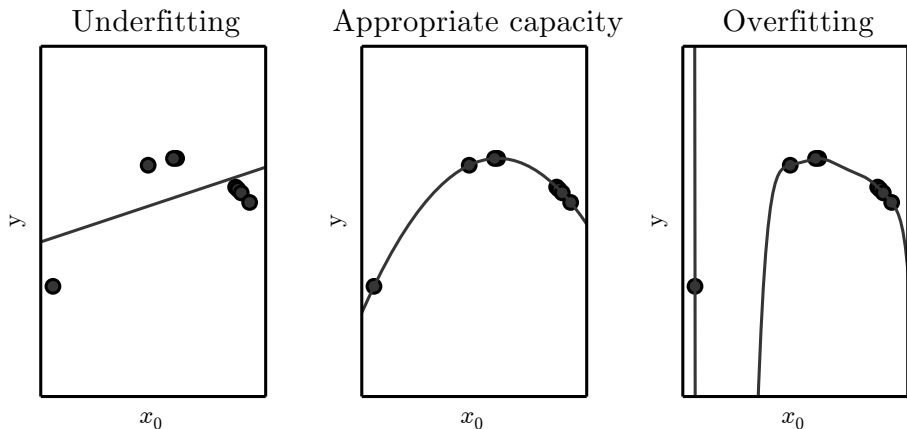


Optimization for nonlinear objective functions.

# Underfitting and Overfitting in Polynomial Estimation

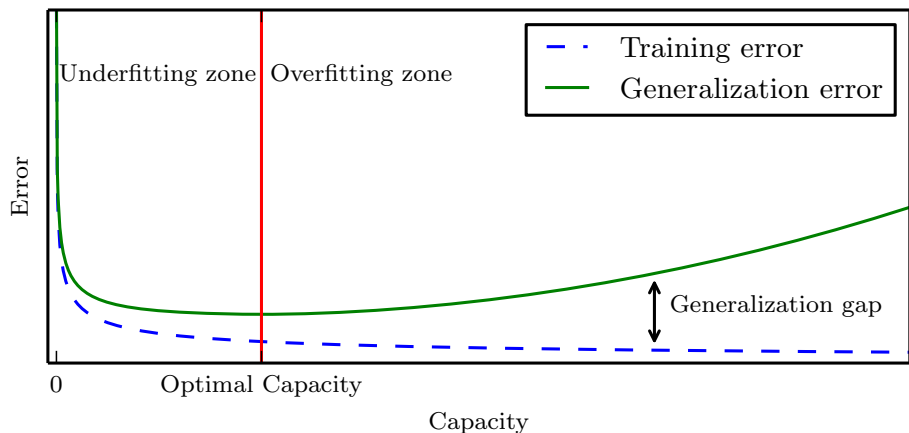
## Model capacity

How flexible a model is.



**Figure:** Left: model too simple (underfitting); Middle: the right model; Right: model too complex (overfitting).

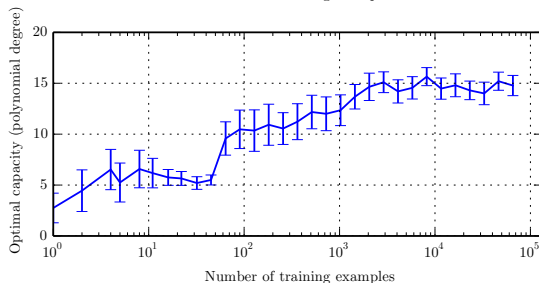
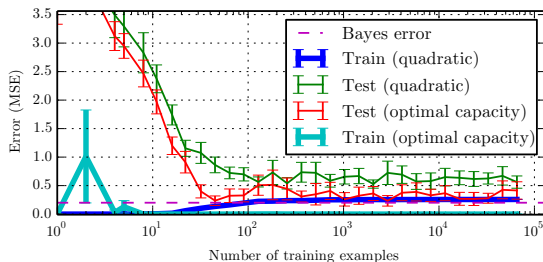
# Generalization and Capacity



**Figure:** Training error would decrease with increasing model complexity; whereas the generalization error (applied on unseen data) achieves its lowest error on the right model complexity.

# Train Set Size VS. Error

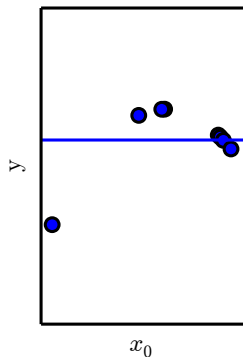
- Bayes error: the best possible error from theory.
- Top: With quadratic capacity: 1) training error increases because more data is harder to training; test error keeps decreasing because it gets more information. 2) does not have enough capacity to make the test error lower than the Bayes error.
- Bottom: As the training set size increases, the optimal capacity increases.



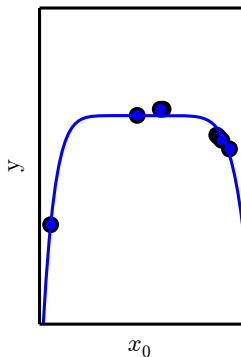
# Avoid Overfitting by Weight Decay

$$J(\mathbf{w}) \triangleq \text{MSE}_{\text{train}} + \lambda \mathbf{w}^T \mathbf{w}$$

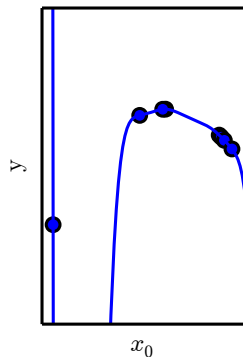
Underfitting  
(Excessive  $\lambda$ )



Appropriate weight decay  
(Medium  $\lambda$ )

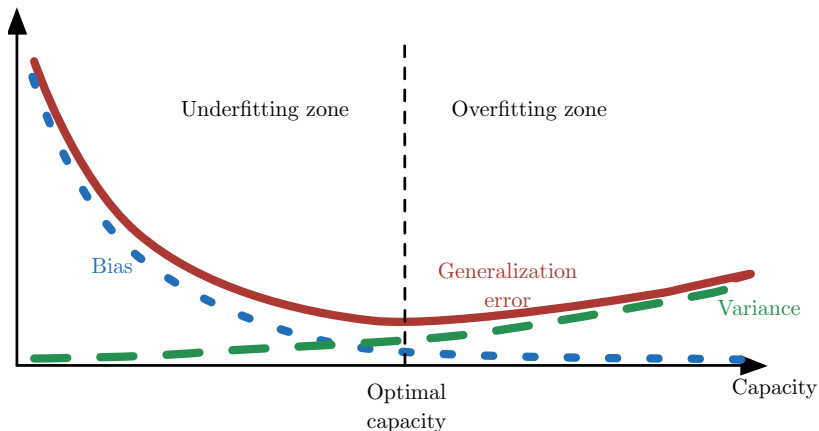


Overfitting  
( $\lambda \rightarrow 0$ )



## Bias and Variance

- In the underfitting zone, we usually get high bias (model doesn't fitting the data), but low variance (model behaviors similarly for different test data).
- Opposite in the overfitting zone.

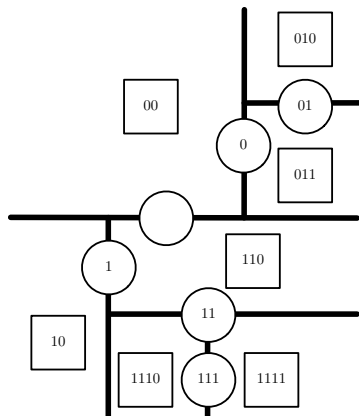
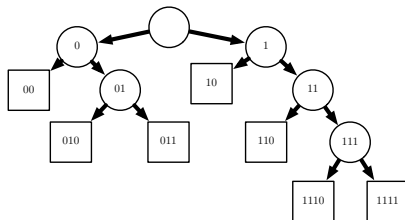




## Traditional Nonlinear Models: Decision Tree

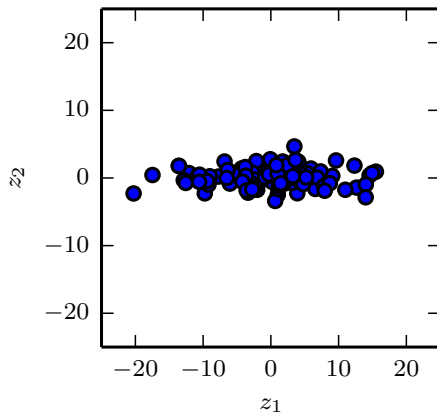
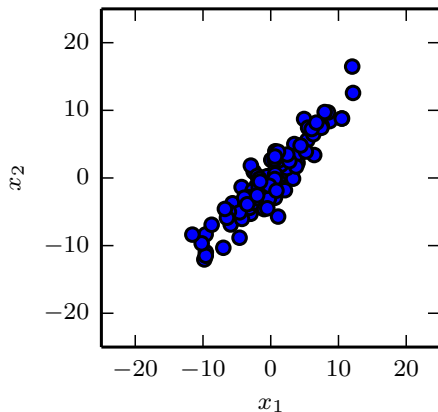
In each step, choose a feature, and separate the data according to the chosen feature:

- Hand-designed features; hard to choose appropriate feature in each step.



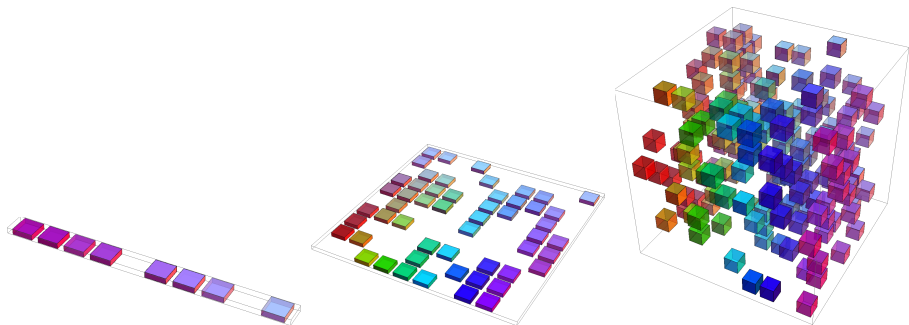
# Principal Component Analysis

Find a direction that has the most variance (most informative) (right) from the original data (left)



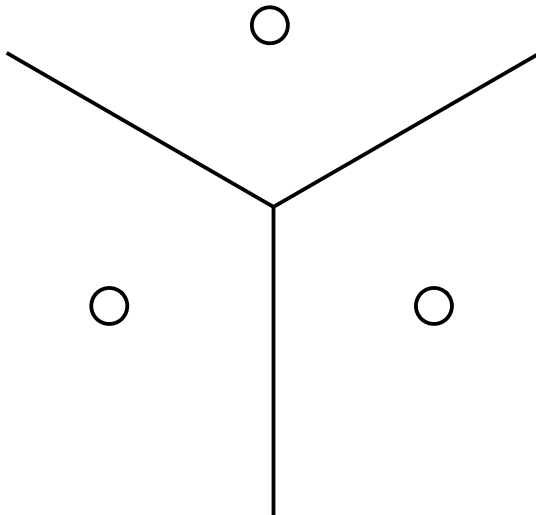
## Curse of Dimensionality

- The increase of data can not meet the increase of model complexity: different configurations of model parameters grows exponentially w.r.t. the dimension, thus some regions of the model parameter space does not have data associated.
- Or data is typically sparse in high dimensional space.



## Nearest Neighbor

- Classify a test data to the class associated with its nearest train data.



## Manifold Learning

- Data typically not flat in the original space.
- They lie in a low-dimensional manifold.
- Fit for deep learning since the manifold is typically nonlinear.

