UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Fall 2024)

Machine Learning

Lecture 10: Regularization

Chong Liu

Assistant Professor of Computer Science

Oct 1, 2024

Announcement

- Homework 1 is due at midnight!
  - Late submissions before tomorrow midnight receive only **half of earned** credits.
  - Submissions after tomorrow midnight receive **NO** credits.
- Homework 2 will be released later today.

Recap: Linear regression

- Stochastic Gradient Descent (SGD)
  - Using a stochastic approximation of the gradient:

$$\theta_{t+1} = \theta_t - \eta_t \hat{\nabla} f(\theta_t)$$

  - Calculated by one data point randomly sampled from dataset
- Linear regression
  - $\hat{w} = \text{argmin}_w \frac{1}{n} \sum_{i=1}^{n} (x_i^T w - y_i)^2 = \text{argmin}_w \|Xw - y\|_2^2$
  - Direct solver: $\hat{w} = (X^T X)^{-1} X^T y$
  - GD: $w \leftarrow w - 2\eta X^T (Xw - y)$, time complexity $O(ndT)$
  - SGD: $w \leftarrow w - 2\eta {x_i}^T ({x_i}^T w - y_i)$, time complexity $O(dT)$

Recap: Two problems of supervised learning

	Classification		Regression
	Binary classification	Multi-class classification	
Feature space	$\mathbb{R}^d$	$\mathbb{R}^d$	$\mathbb{R}^d$
Label space	{-1, 1}	{1, 2, 3, …, K}	$\mathbb{R}$
Performance metric	Classification error (0-1 loss) for test data	Classification error (0-1 loss) for test data	Mean Square Error
Popular surrogate loss (for training)	Logistic loss / exponential loss / square loss	Multiclass logistic loss (Cross-Entropy loss)	Square loss

Today

- Linear regression in curve fitting
- Problem of overfitting
- Regularization prevents overfitting

Example of regression - Curve fitting: how to train a function fitting blue dots?

- Input:

$$x \in \mathcal{X} = [0, 1] \subset \mathbb{R}$$

- Output:

$$y \in \mathcal{Y} = \mathbb{R}$$

- Data:

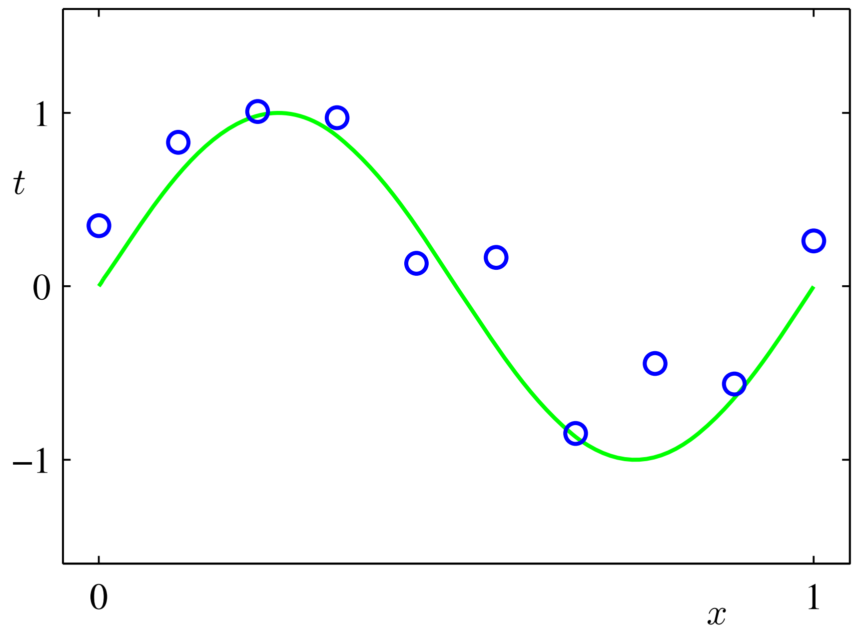
$$(x_1, y_1), \dots, (x_n, y_n)$$

- Ground truth:

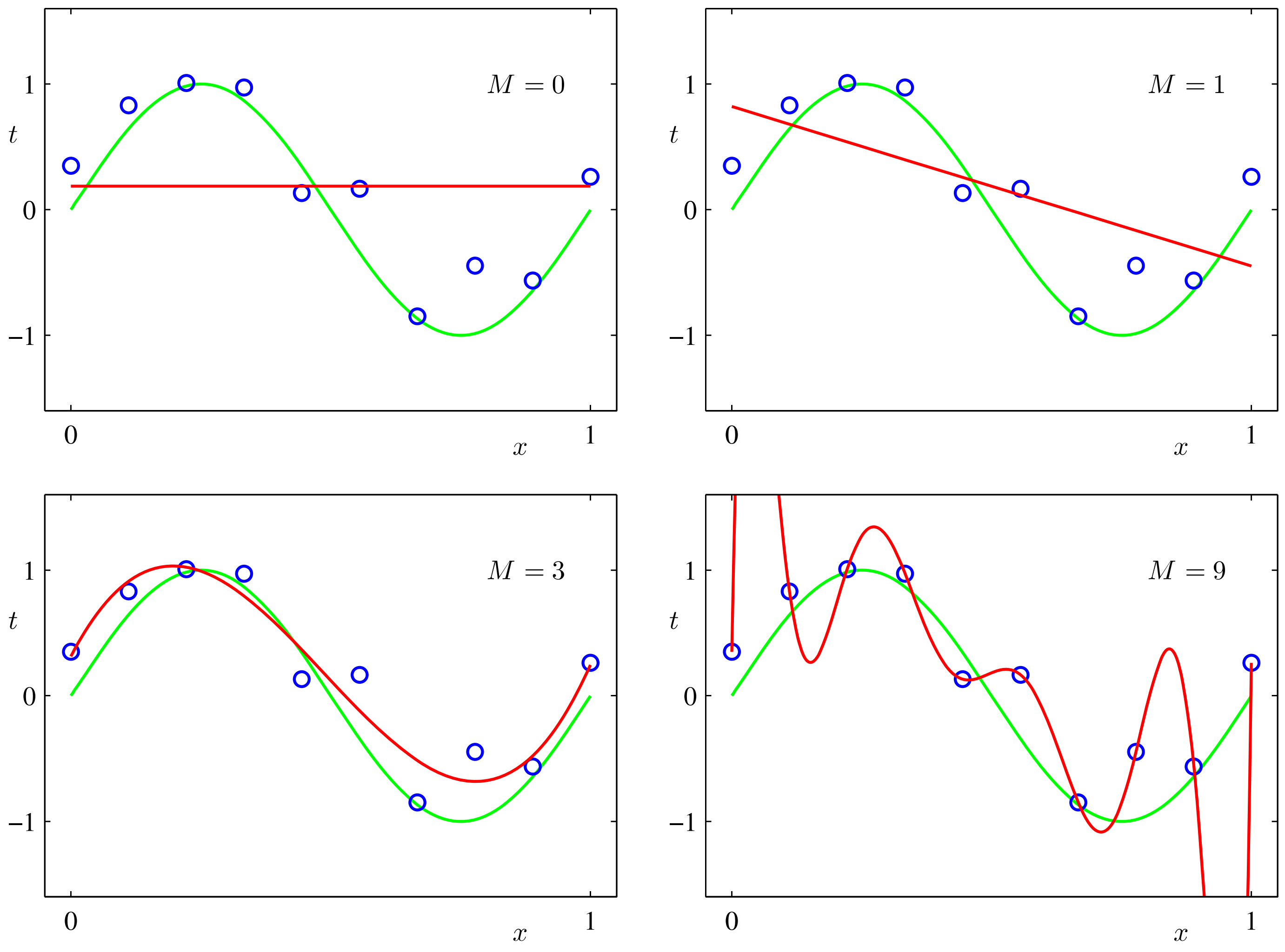
$$f_0(x) = \sin(2\pi x)$$

- Hypothesis?

$$f(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$$

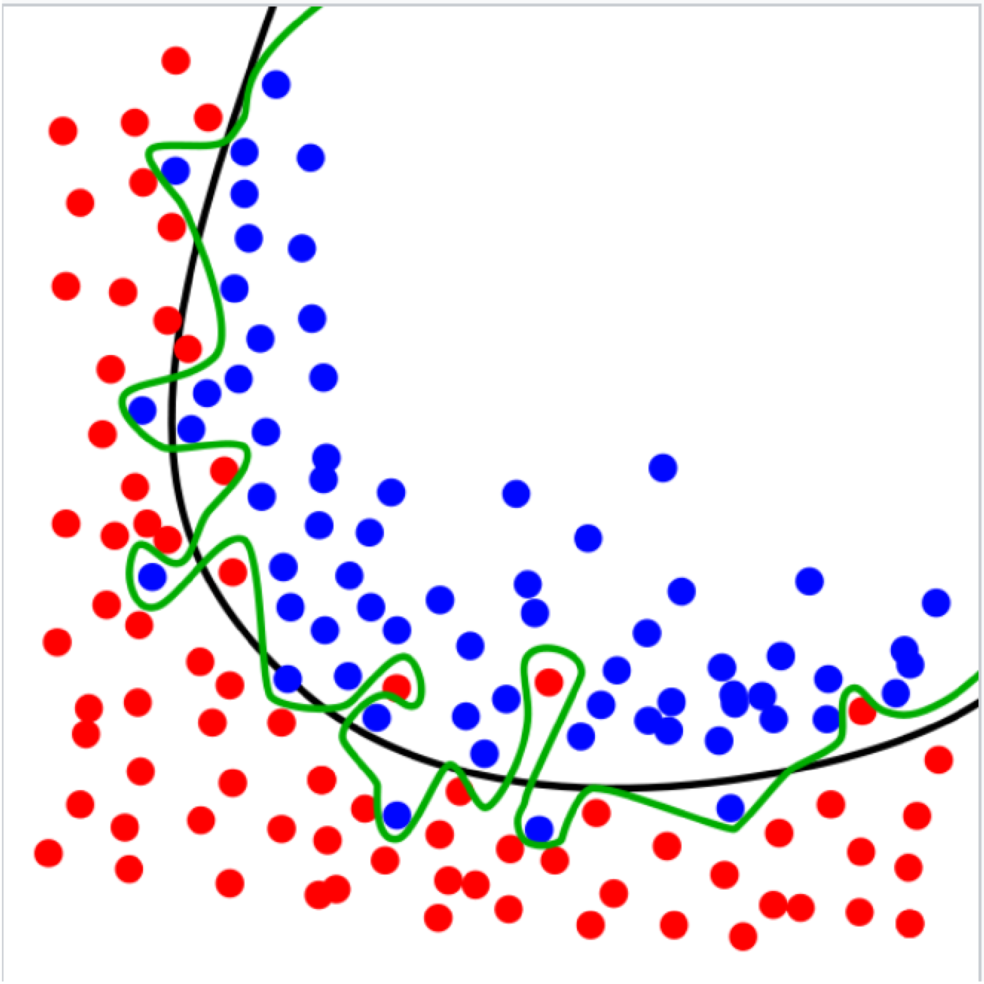


Polynomial regression under square loss



Problem of overfitting!

Recap: The problem of Overfitting



The green line represents an overfitted model.

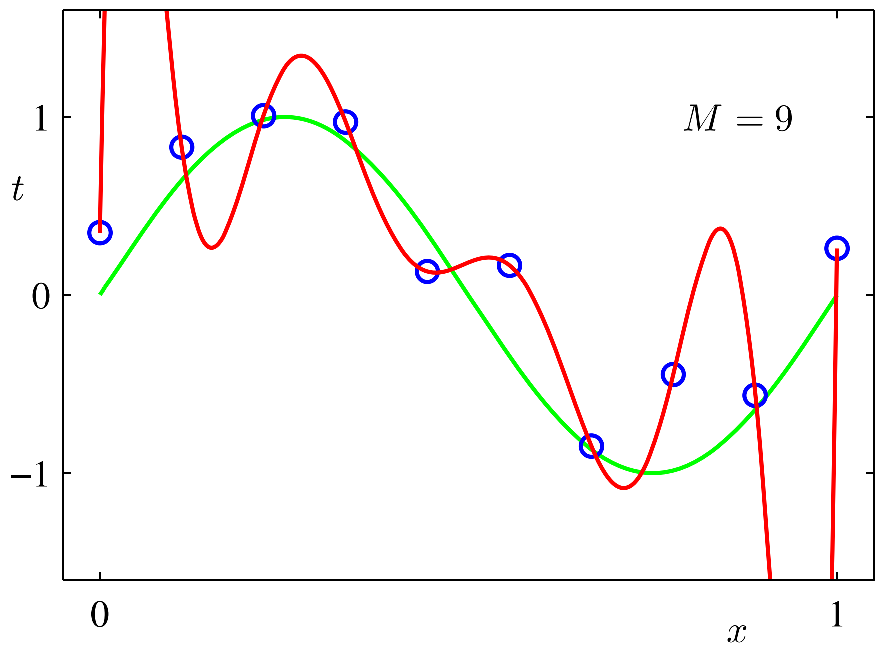
1. Best follows the training data
2. Too dependent on training data
3. More likely to fail (higher error rate) than black line on new unseen test data

Discussion: examples of overfitting in our learning as human beings?

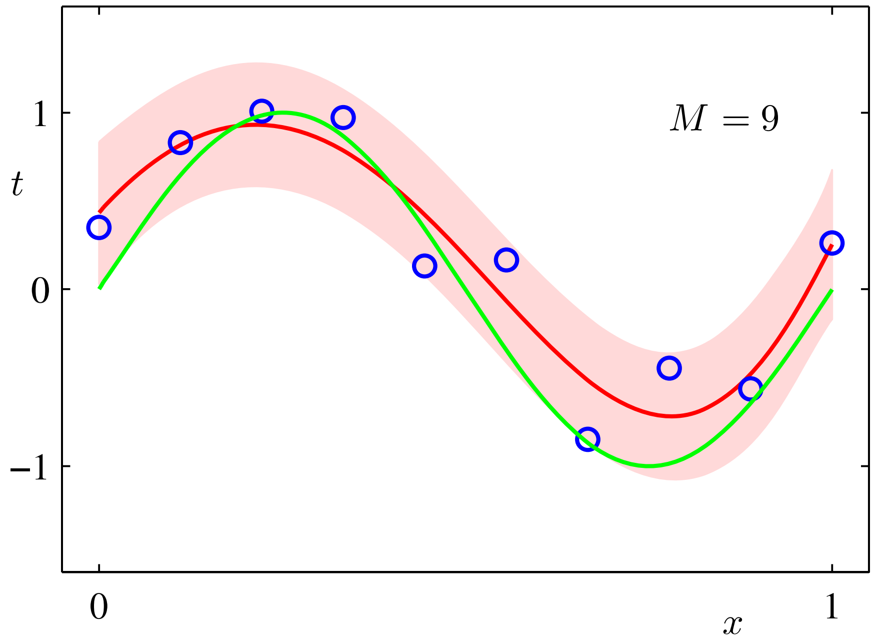
Regularization prevents overfitting!

- Same model:

$$f(x) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$$



No regularization



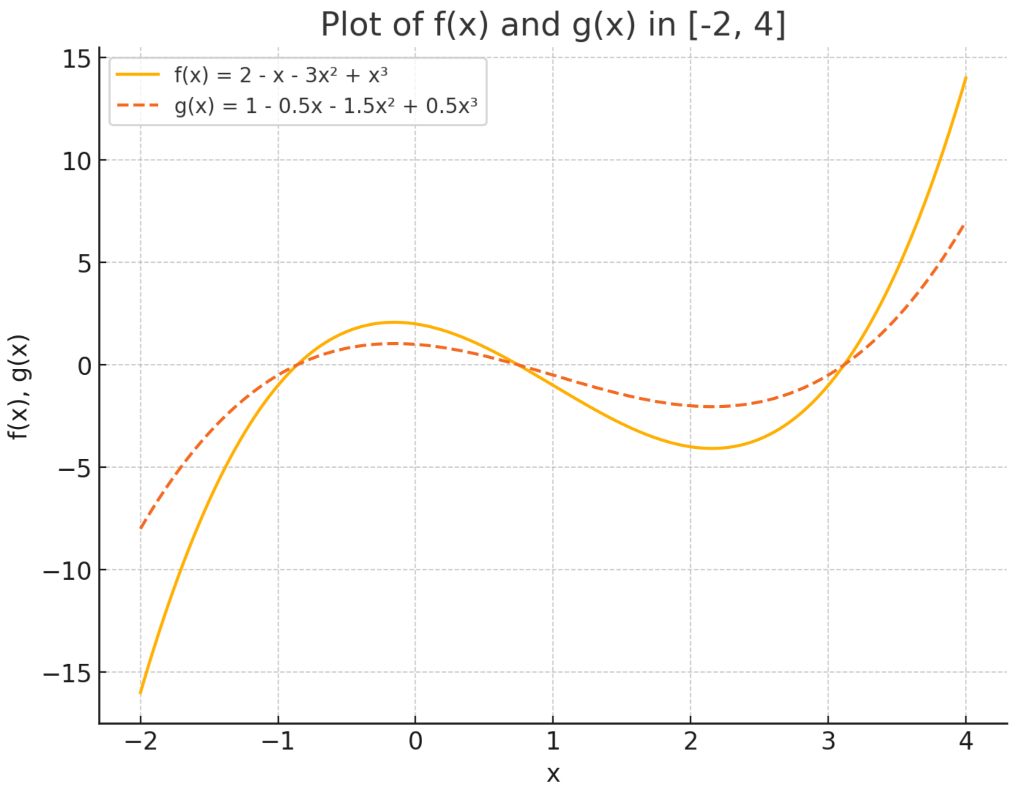
With regularization

How does regularization work?

- Regularization controls the parameter complexity (norms of parameter).
  - p-norm regularized least square

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n}\|X\theta - y\|_2^2 + \lambda\|\theta\|_p^p$$

- In-class exercise:
  - Find L-2 norm of $x = [2, -1, -3, 1]$ and $x' = [1, -0.5, -1.5, 0.5]$
  - Plot $f(x) = 2 - x - 3x^2 + x^3$ and $g(x) = 1 - 0.5x - 1.5x^2 + 0.5x^3$ in $[-2, 4]$



Different choices of regularization

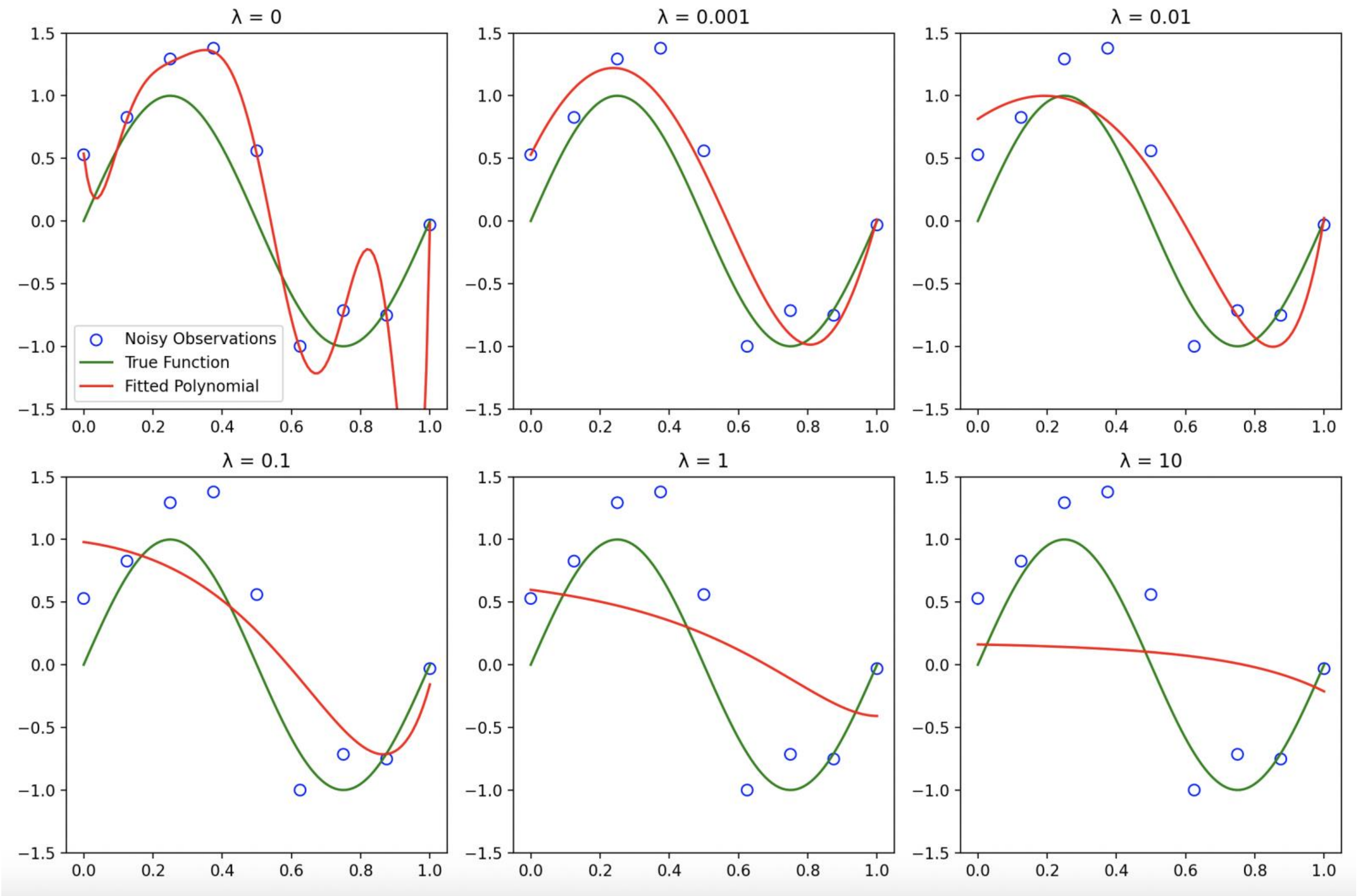
- p-norm regularized least square

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \|X\theta - y\|_2^2 + \lambda \|\theta\|_p^p$$

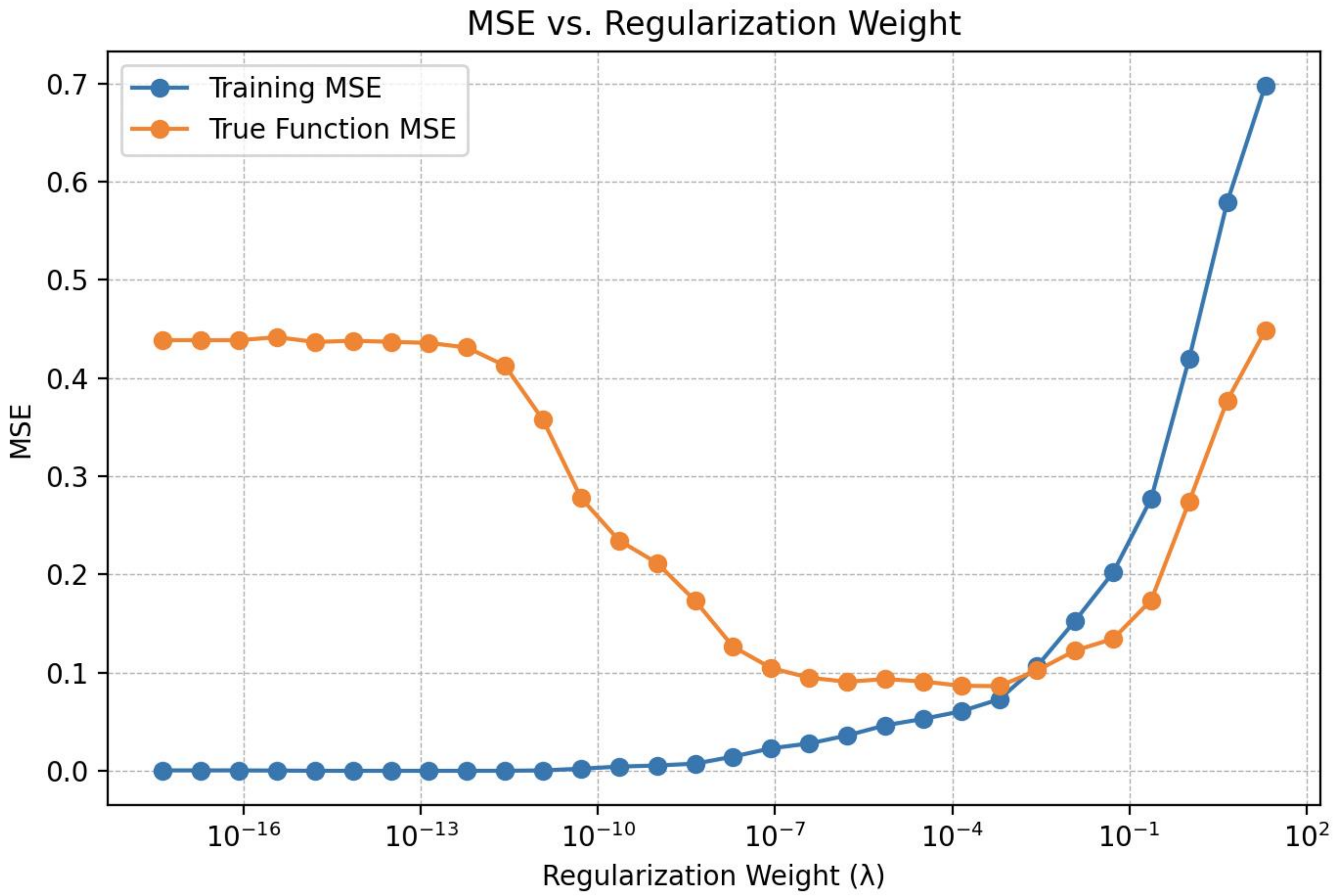
  - when p=2, this is called "Ridge Regression"
  - when p=1, this is called "Lasso"

- In-class exercise: work out the direct solver for Ridge Regression.

Fitted curve as L-2 regularization weight increases



The mean square errors as we adjust the L2 regularization weight



- Why is test error in a U-shaped curve?
  - Due to bias-variance tradeoff (after midterm)
  - Regularization weight cannot be too large or too small

- Discussion:

  Why training error increases with regularization weight?

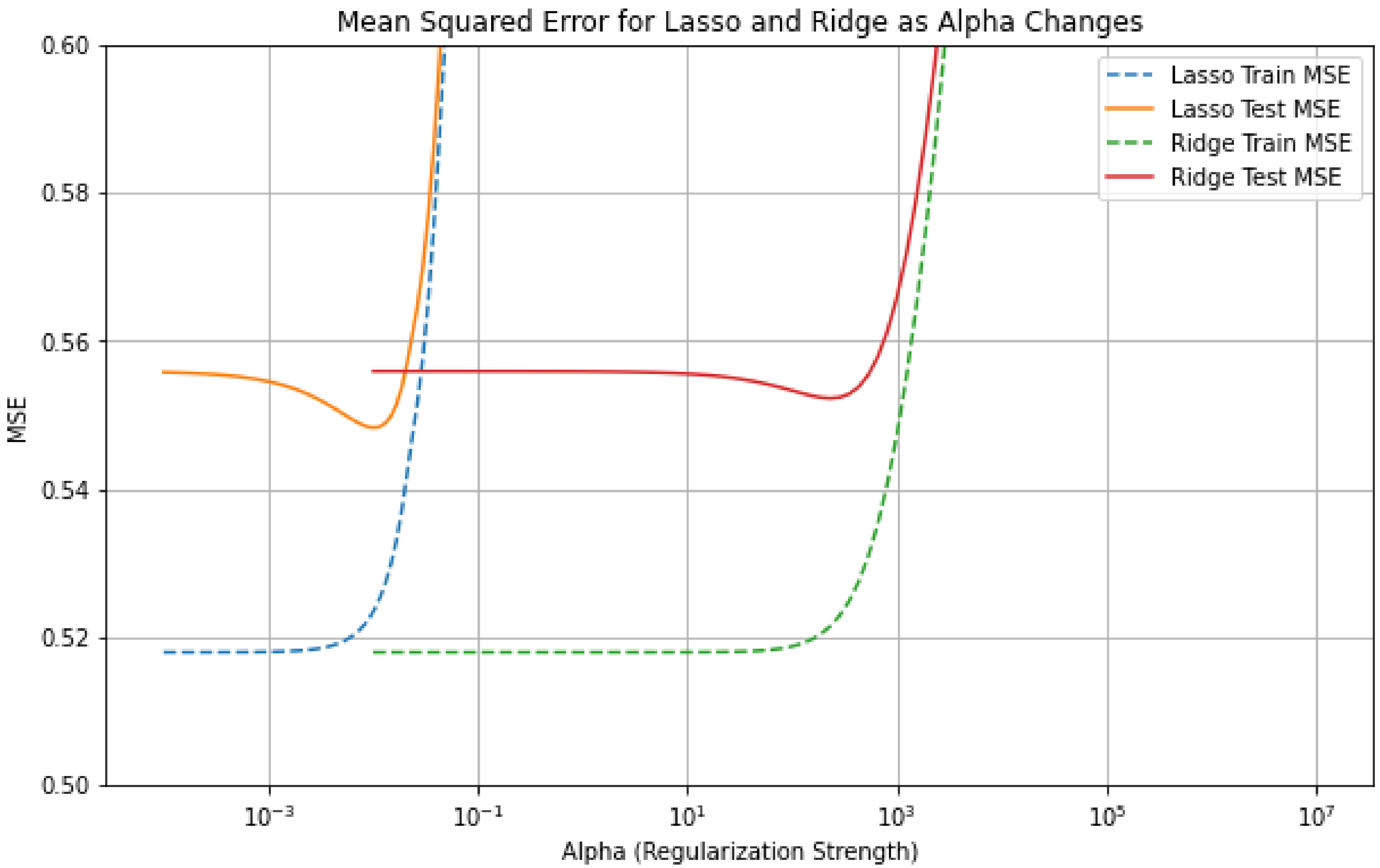
Case study: Housing price dataset

- Example data

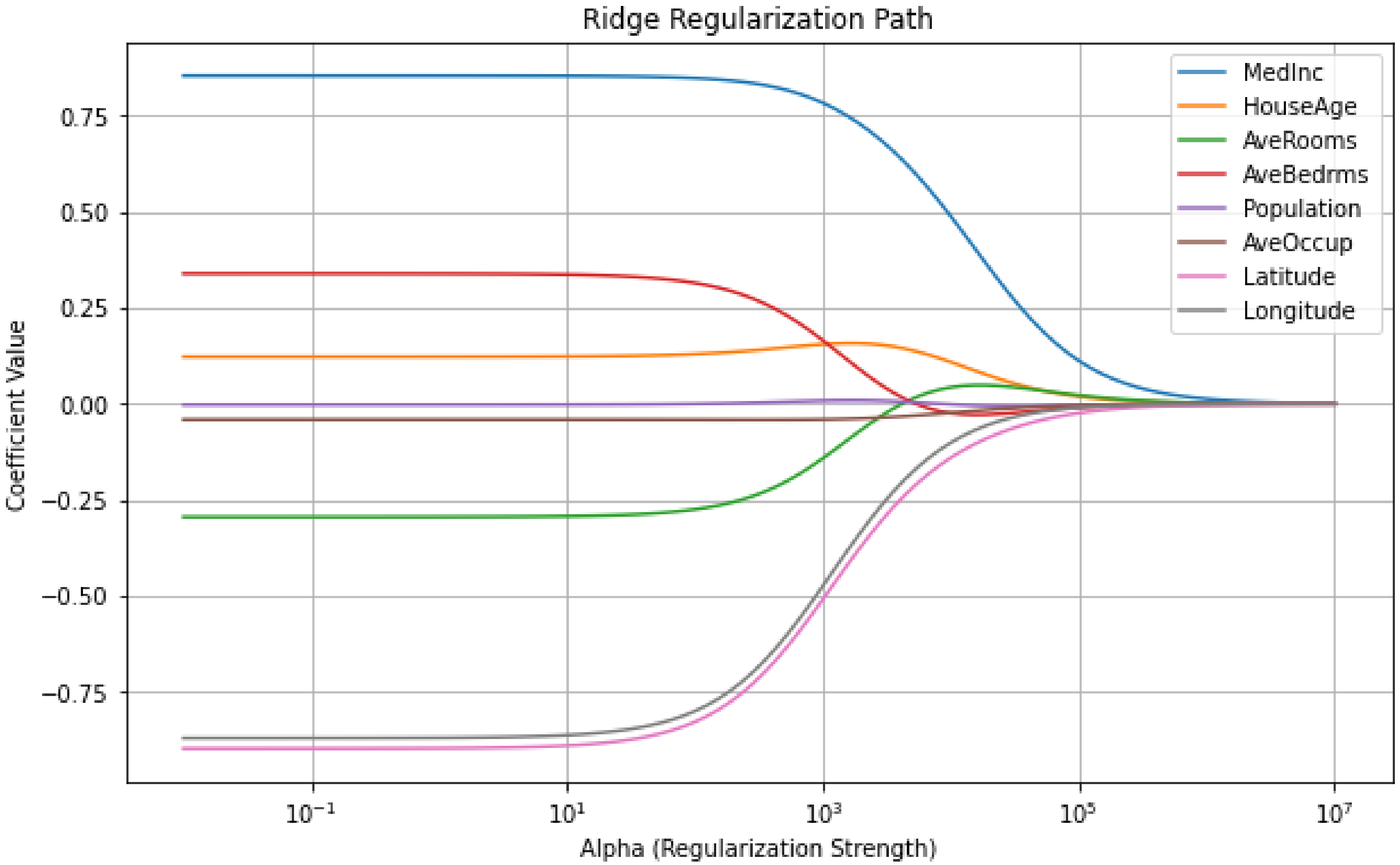
	MedInc	HouseAge	AveRooms	AveBedrms	Population	AveOccup	Latitude	Longitude	Target
0	8.3252	41.0	6.984127	1.023810	322.0	2.555556	37.88	-122.23	4.526
1	8.3014	21.0	6.238137	0.971880	2401.0	2.109842	37.86	-122.22	3.585
2	7.2574	52.0	8.288136	1.073446	496.0	2.802260	37.85	-122.24	3.521
3	5.6431	52.0	5.817352	1.073059	558.0	2.547945	37.85	-122.25	3.413
4	3.8462	52.0	6.281853	1.081081	565.0	2.181467	37.85	-122.25	3.422

- Questions one can ask:
  - How well can one use the 8 features to predict house price (i.e., Target)?
  - Which feature is more predictive with house price?
  - What is the effect of regularization?

The MSE vs regularization weight



The “Regularization path” for L2-regularization



How to interpret the fitted coefficients?

- The “sign” indicates positive or negative correlation with the label
- The “magnitude” indicates how strongly correlated.

Summary

- Linear regression
  - Solving the Least Square problem {with GD, SGD and direct solver}
- Regularization
  - Controls the parameter complexity of the fitted function
  - Prevents overfitting!
  - Different regularization: L-2 (most popular) and L-1
- Case study: Predict Housing Price
  - Effect of regularization on training test and test error
  - Regularization path (Effect of regularization on coefficients)