Lecture 05. Regularization

Xin Chen

These slides are based on slides from Mahdi Roozbahani

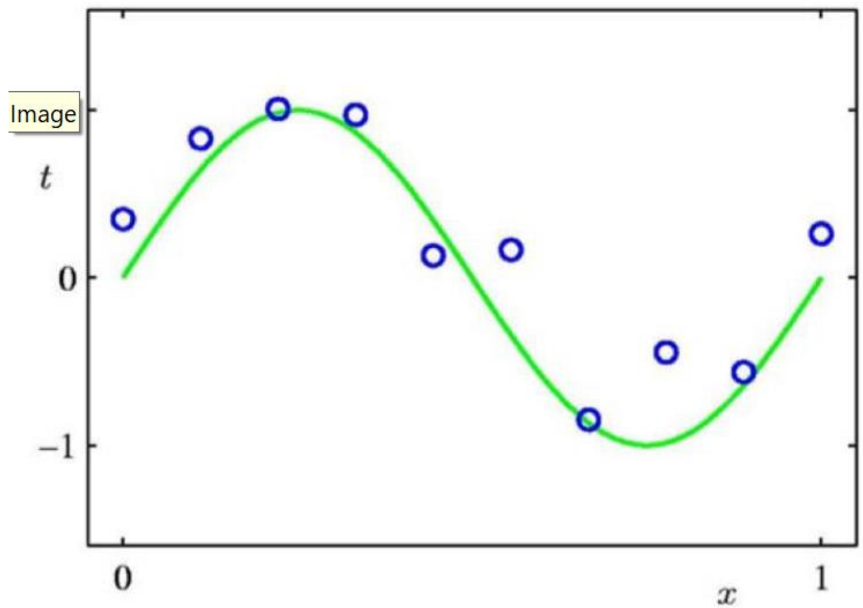
Logistics

- Form your project team
- Schedule of assignments and project
  - Every two weeks, there will be a new homework. In total, we have 4.
  - Project schedule:
    - Next Wednesday **(Jun 3rd)** our lecture will be about the project requirement.
    - This weekend, I will share some dataset that you may use for your project. I will create a excel file that briefly introduces your project.
    - Form your team by the end of next week and I will assign you a team randomly on Friday **Jun 5th**.
    - Project proposal is due on Sun **Jun 14th.**
    - Project presentation is on Wed **July 13th**.

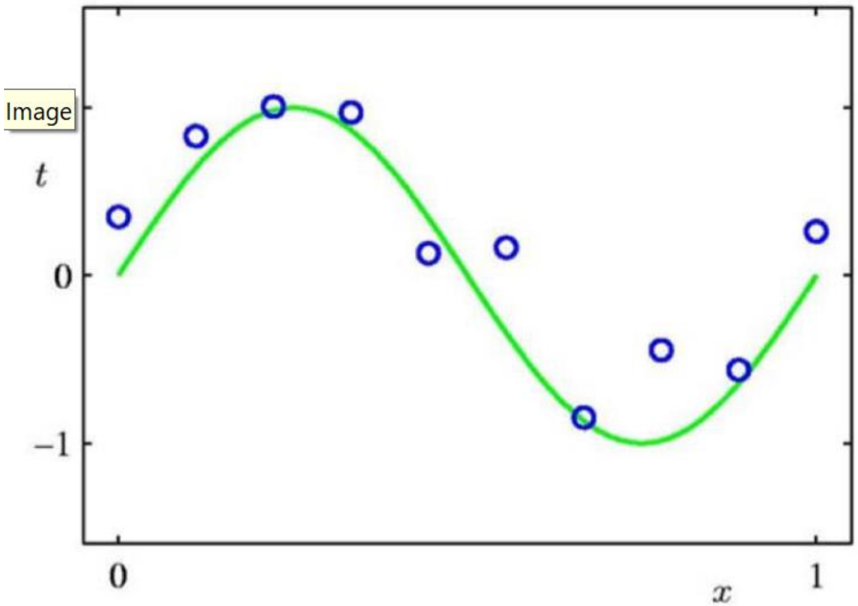
Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization length

Regression



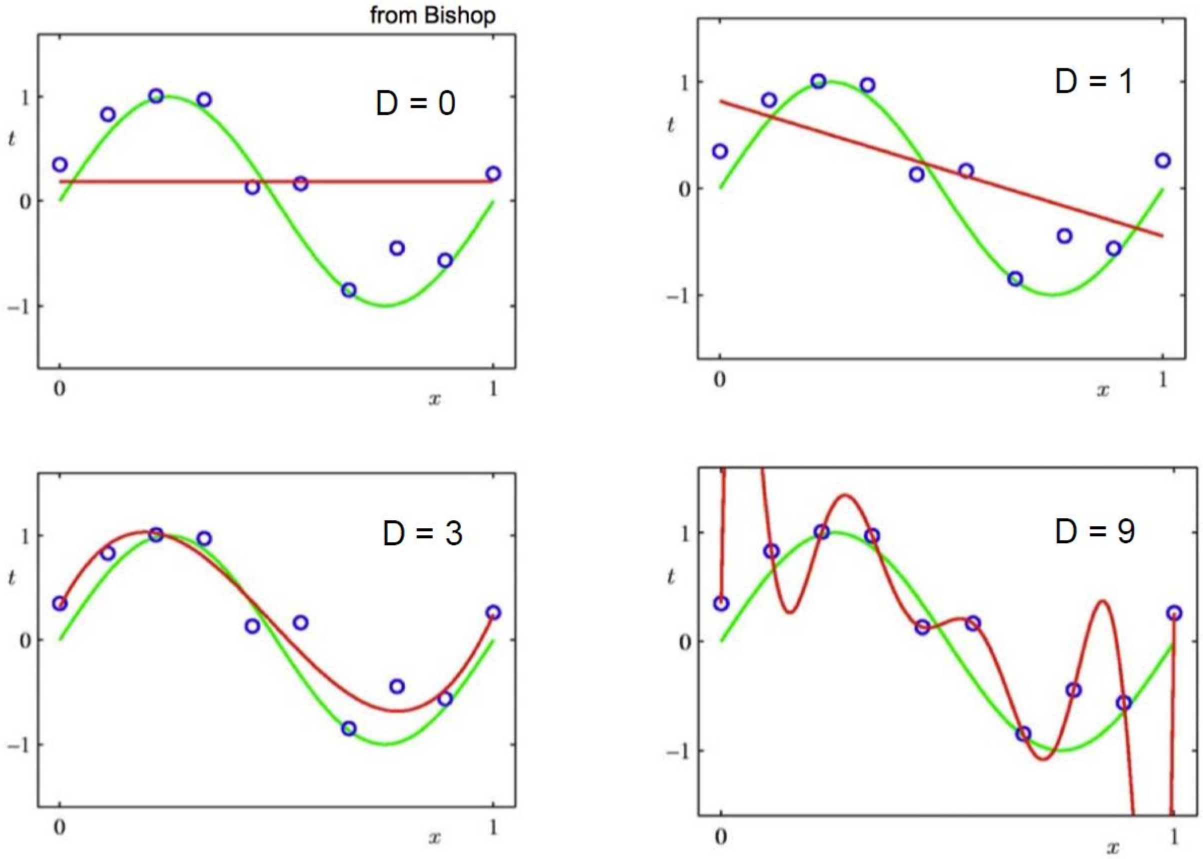
- Suppose we are given a training set of N observations $\{(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n)\}$
- Regression problem is to estimate $y(x)$ from the dataset.



- Want to fit this data to a polynomial regression model: $y = \theta_0 + \theta_1 x^1 + \ldots + \theta_d x^d + \epsilon$
- Let $z = \{1, x^1, x^2, \ldots x^d\} \in R^d$ and $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$

$$\rightarrow y = z\theta$$

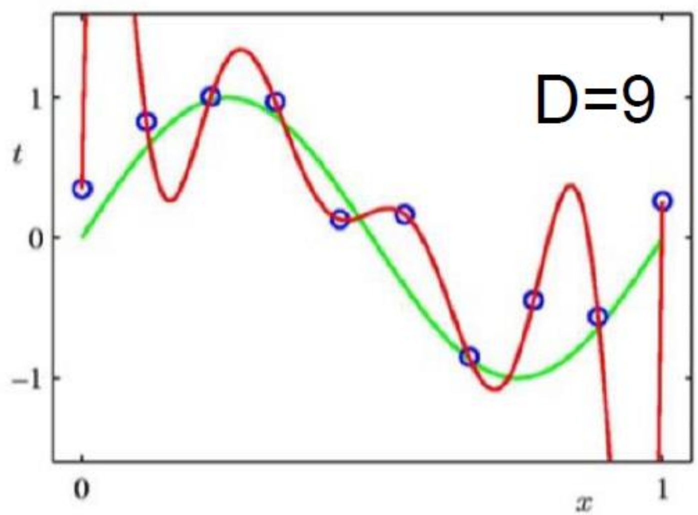
Which one is better?

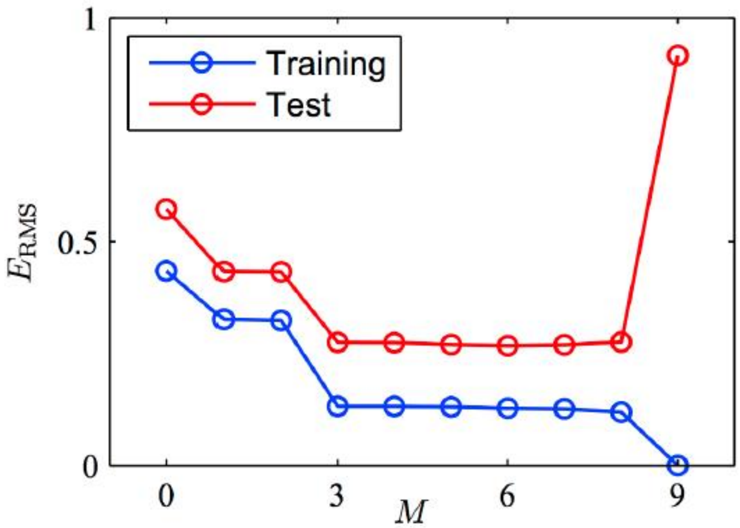


Can we increase the maximal polynomial degree to a very large dimension, as a “safe” solution?

– No, this can lead to overfitting !!!

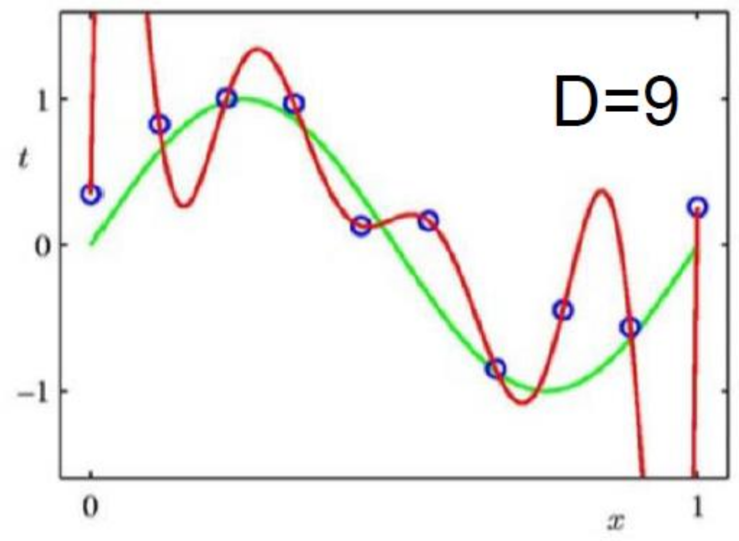
The overfitting problem





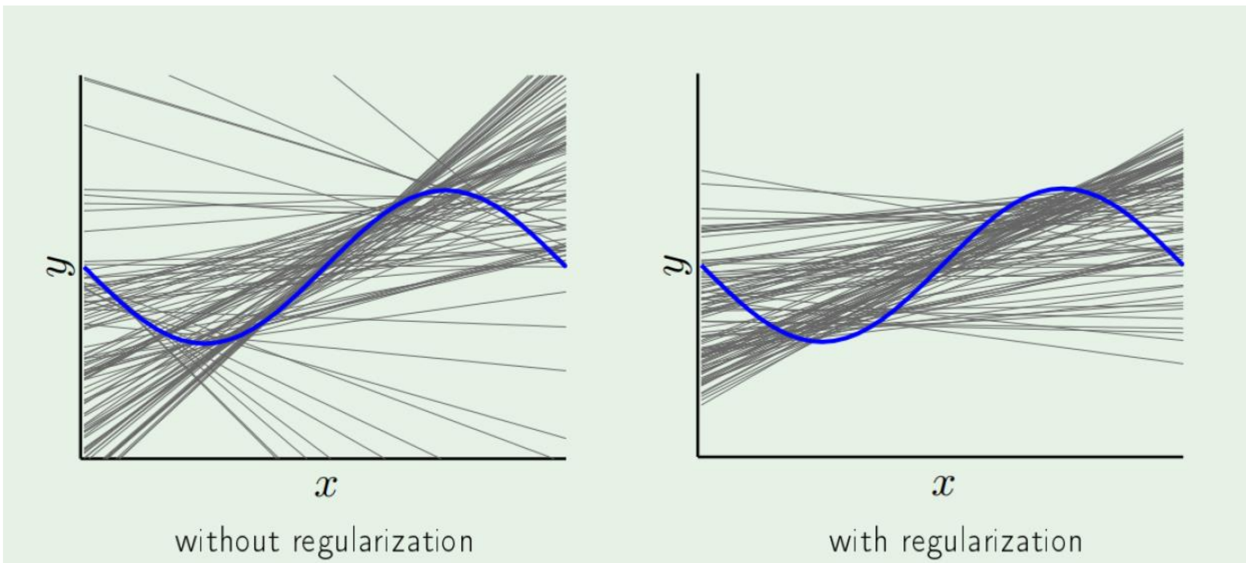
- The training error is very low, but the error on test set is large.
- The model captures not only patterns but also noisy nuisances in the training data.

The overfitting problem



- In regression, overfitting is often associated with large weights (severe oscillation).
- How can we address overfitting?

Regularization (smart way to cure overfitting disease)

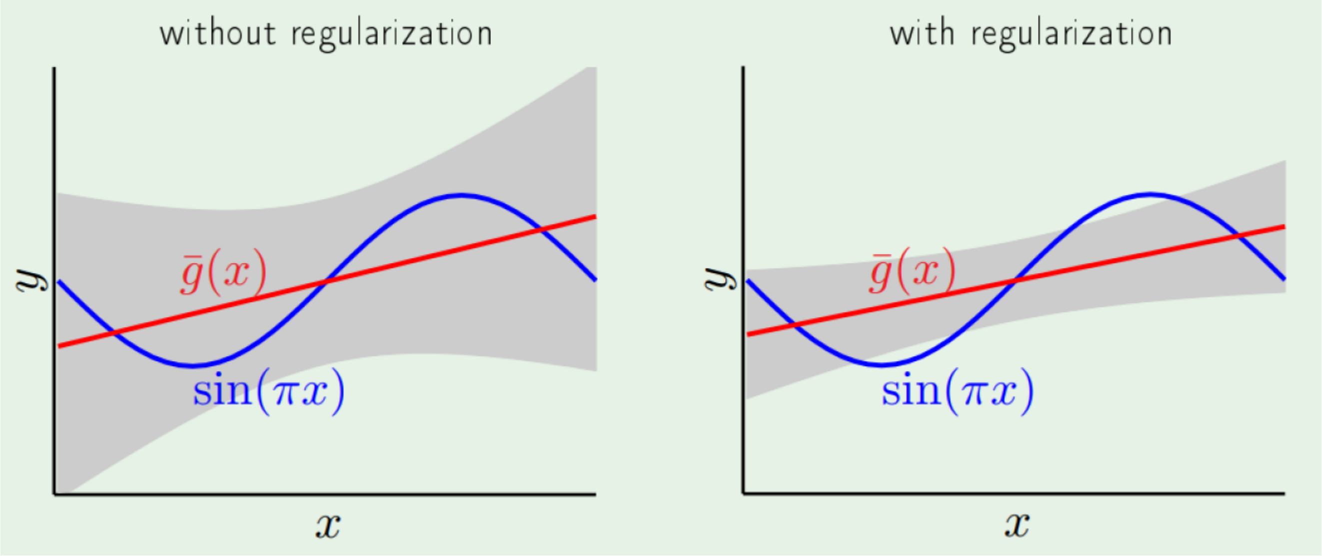


Put a break on fitting

- Fit a linear line on sinusoidal with just two points.

Who is the winner?

$\bar{g}(x)$ is the average over all lines



Bias=0.21; var=1.69

Bias=0.23; var=0.33

Regularized learning

Minimize $E(\theta) + \frac{\lambda}{N}\theta^T\theta$

Why this term leads to regularization of parameters?

Cost function: squared loss

$$E(\theta) = \frac{1}{N}\underbrace{\sum_{i=1}^{N}\{f(x_i,\theta)\}^2}_{\text{Loss function}} + \underbrace{\frac{\lambda}{N}\theta^T\theta}_{\text{Regularization}}$$

Regularization is just constraining the weights(θ)

- Want to fit this data to a polynomial regression model: $y = \theta_0 + \theta_1 x^1 + \ldots + \theta_d x^d + \epsilon$
- Let $z = \{1, x^1, x^2, \ldots x^d\} \in R^d$ and $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$

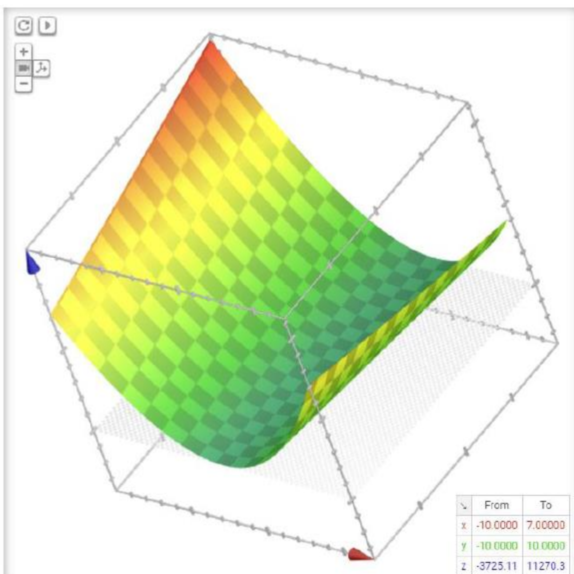
$$\begin{cases} \text{Minimize } \mathrm{E}(\theta) = \frac{1}{N}(Z\theta - y)^T(Z\theta - y) \\ \text{Subject to } \theta^t\theta \leq C \end{cases}$$

- For simplicity: let's call θ_{lin} as weights' solution for non-constrained one and θ for the constraint model.

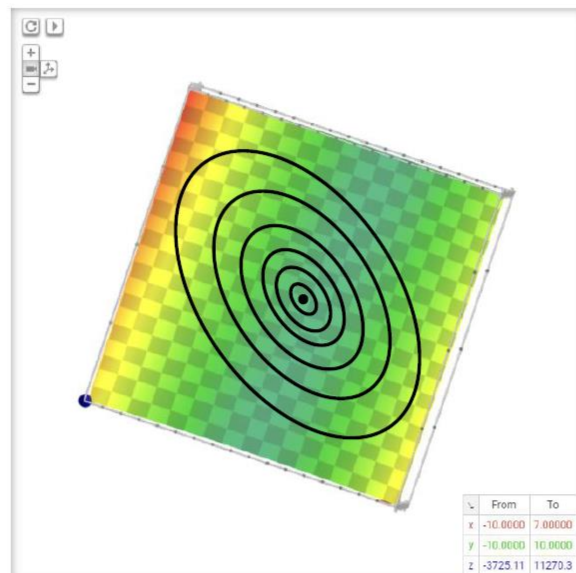
Consider an example

Let d=2: $y = \theta_0 + \theta_1 Z_1 + \theta_2 Z_2$

An example: $\mathrm{E}(\theta) = ([5 + 10x] - y)^2$



3D view

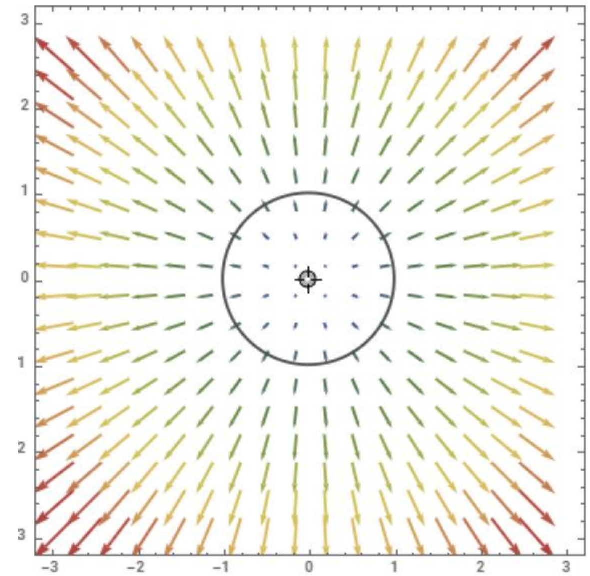


Top view

Gradient $\theta^T\theta$

$$\theta = \begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix} \qquad \Rightarrow \theta^t\,\theta = \theta_0^2 + \theta_1^2$$

$$\nabla(\theta^T\theta) = \begin{bmatrix}\dfrac{\partial}{\partial(\theta_0)}(\theta^T\theta)\\ \dfrac{\partial}{\partial(\theta_1)}(\theta^T\theta)\end{bmatrix} = \begin{bmatrix}2\theta_0\\ 2\theta_1\end{bmatrix} \approx \begin{bmatrix}\theta_0\\ \theta_1\end{bmatrix}$$

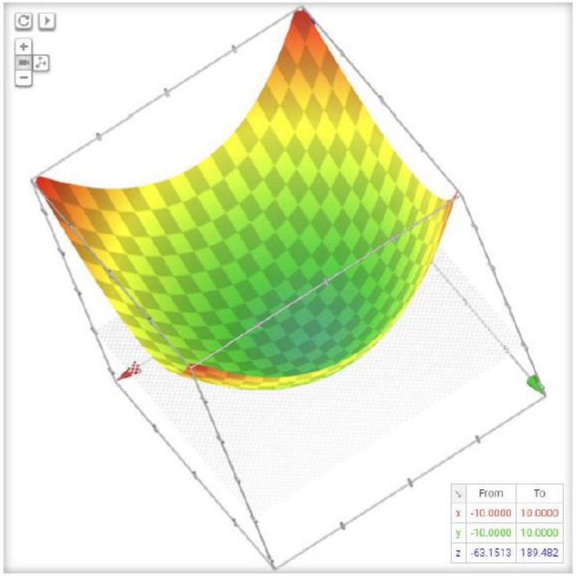


- Imagine you standing at a point (θ_0, θ_1),$\nabla(\theta^T\theta)$ tells you which direction you should go to increase the value of $\theta^T\theta$ most rapidly.

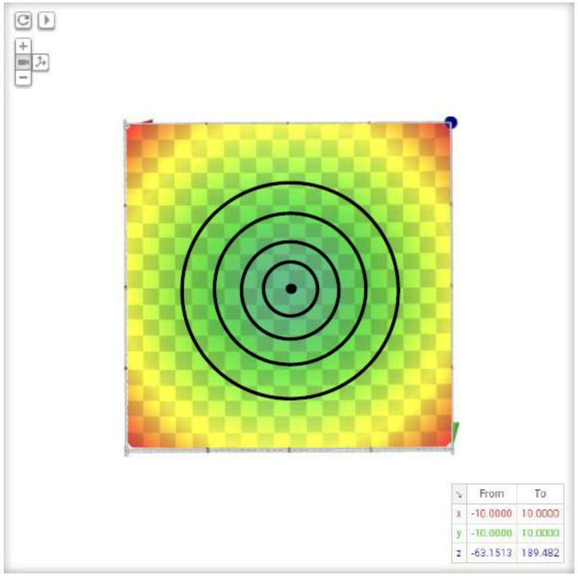
$\nabla(\theta^T\theta)$ is a vector, any line passing through the center of the circle.

Graph of $\theta^T \theta$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \qquad \Rightarrow \theta^t\, \theta = \theta_0^2 + \theta_1^2$$



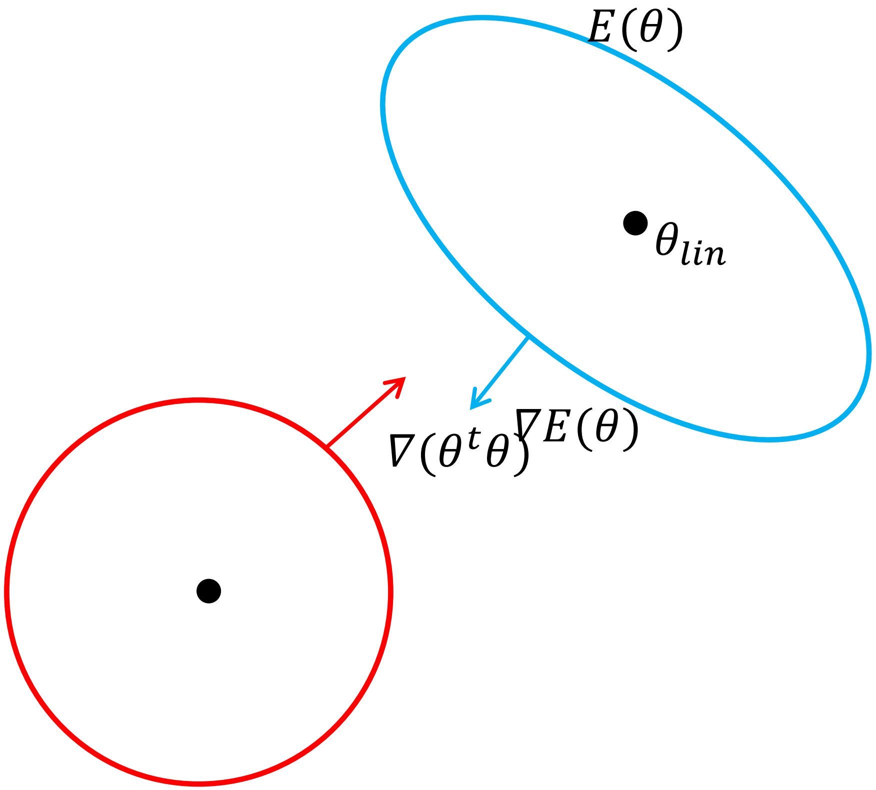
3D view



Top view

$$\begin{cases} \text{Minimize } \mathrm{E}(\theta) = \frac{1}{N}(Z\theta - y)^T(Z\theta - y) \\ \text{Subject to } \theta^t\theta \le C \end{cases}$$

∇E: the gradient (rate) in objective function that minimizes the error (orthogonal to ellipse)

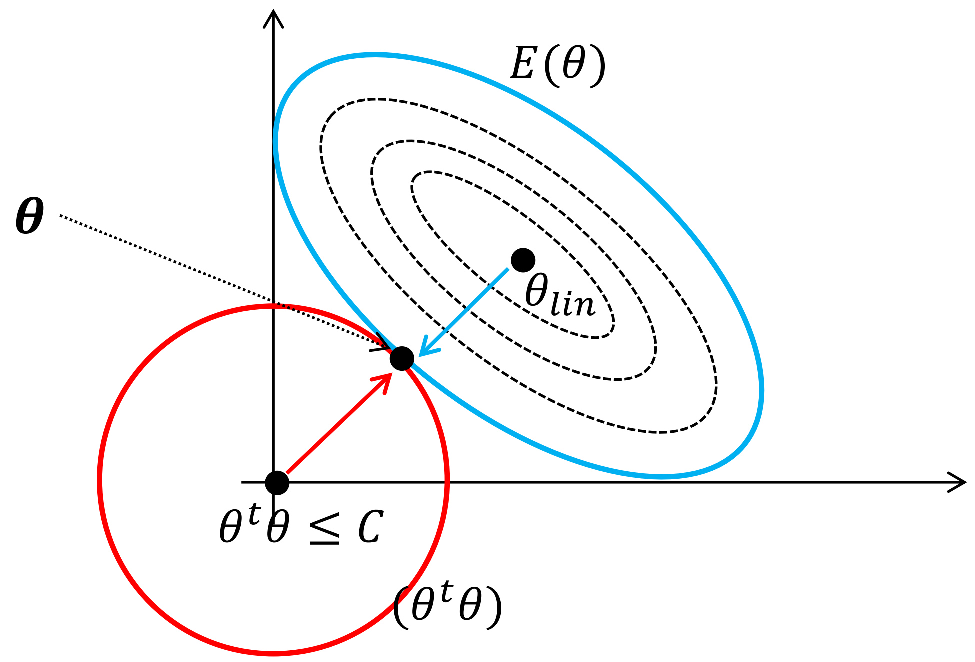


Applying a constraint $\theta^t\theta$, where the best solution happens?

On the boundary of the circle, as it is the closest one to the minimum absolute

Do the integration

Minimize $\mathrm{E}(\theta) + \frac{\lambda}{N}\theta^T\theta$



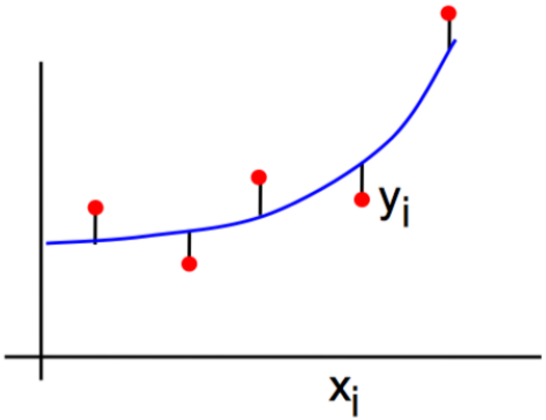
The final solution is $\boldsymbol{\theta}$, after applying the regularization.

Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization length

Ridge Regression

- Cost function-square loss



$$E(\theta) = \frac{1}{N}\sum_{i=1}^{N}\underbrace{\{f(x_i,\theta) - y_i\}^2}_{\text{Loss function}} + \underbrace{\frac{\lambda}{N}||\theta||^2}_{\text{Regularization}}$$

- Regression function for x (1d)

$$y = \ \theta_0 + \theta_1 Z_1 + \ldots + \theta_d Z_d + \epsilon$$

Solving for the weights θ

Write the target and the regressed values as vectors

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{pmatrix} \quad \mathbf{f} = \begin{pmatrix} z(x_1)\theta \\ z(x_2)\theta \\ . \\ . \\ z(x_n)\theta \end{pmatrix} = z\theta = \begin{bmatrix} 1 & z_1(x_1) & \dots & z_d(x_1) \\ 1 & z_1(x_2) & \dots & z_d(x_2) \\ . & & & \\ . & & & \\ 1 & z_1(x_n) & \dots & z_d(x_n) \end{bmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ . \\ . \\ \theta_d \end{pmatrix}$$

An example, with polynomial regression with basic functions up to x^2

$$z\theta = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ . & & . \\ . & & . \\ 1 & x_N & x_N^2 \end{bmatrix} \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{pmatrix}$$

$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\{f(x_i,\theta) - y_i\}^2 + \frac{\lambda}{N}||\theta||^2$$

$$E(\theta) = \frac{1}{N}(y - Z\theta)^2 + \frac{\lambda}{N}||\theta||^2$$

Let's compute derivative w.r.t. θ is zero for minimum.

$$\frac{\tilde{E}(\theta\ )}{d\theta} = -z^T(y - z\theta) + \lambda\theta$$

$$(Z^T Z + \lambda I)\theta = Z^T y$$

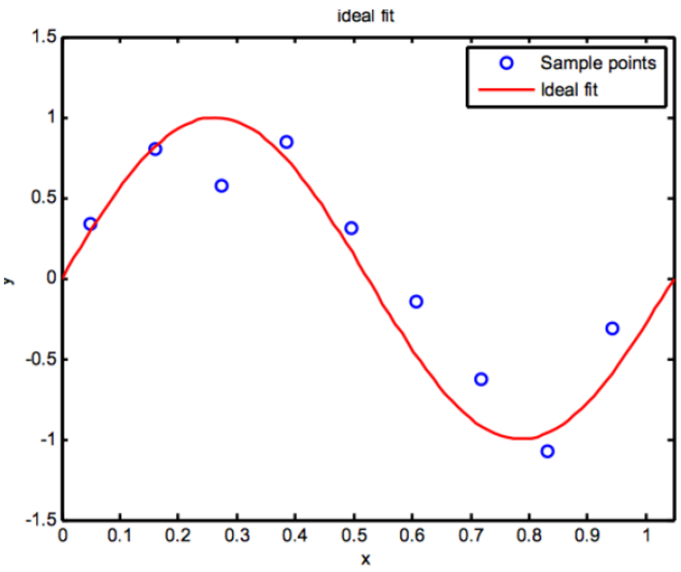
$$\theta = (Z^T Z + \lambda I)^{-1} Z^T y$$

$$\underbrace{\theta}_{D\times 1} = \underbrace{(Z^T Z + \lambda I)^{-1}}_{D\times D} \underbrace{Z^T}_{D\times N} \underbrace{y}_{N \times 1}$$

- If λ =0 (no regularization), then $\theta = (Z^T Z)^{-1} Z^T \mathrm{y}$
- If $\lambda = \infty$, $\theta = \frac{1}{\lambda} Z^T \mathrm{y} \rightarrow 0$
- Adding the term λI improves the conditioning of the inverse, since if Z is not full rank, then $Z^T Z + \lambda I$ will be (for sufficiently large λ).

Ridge Regression Example

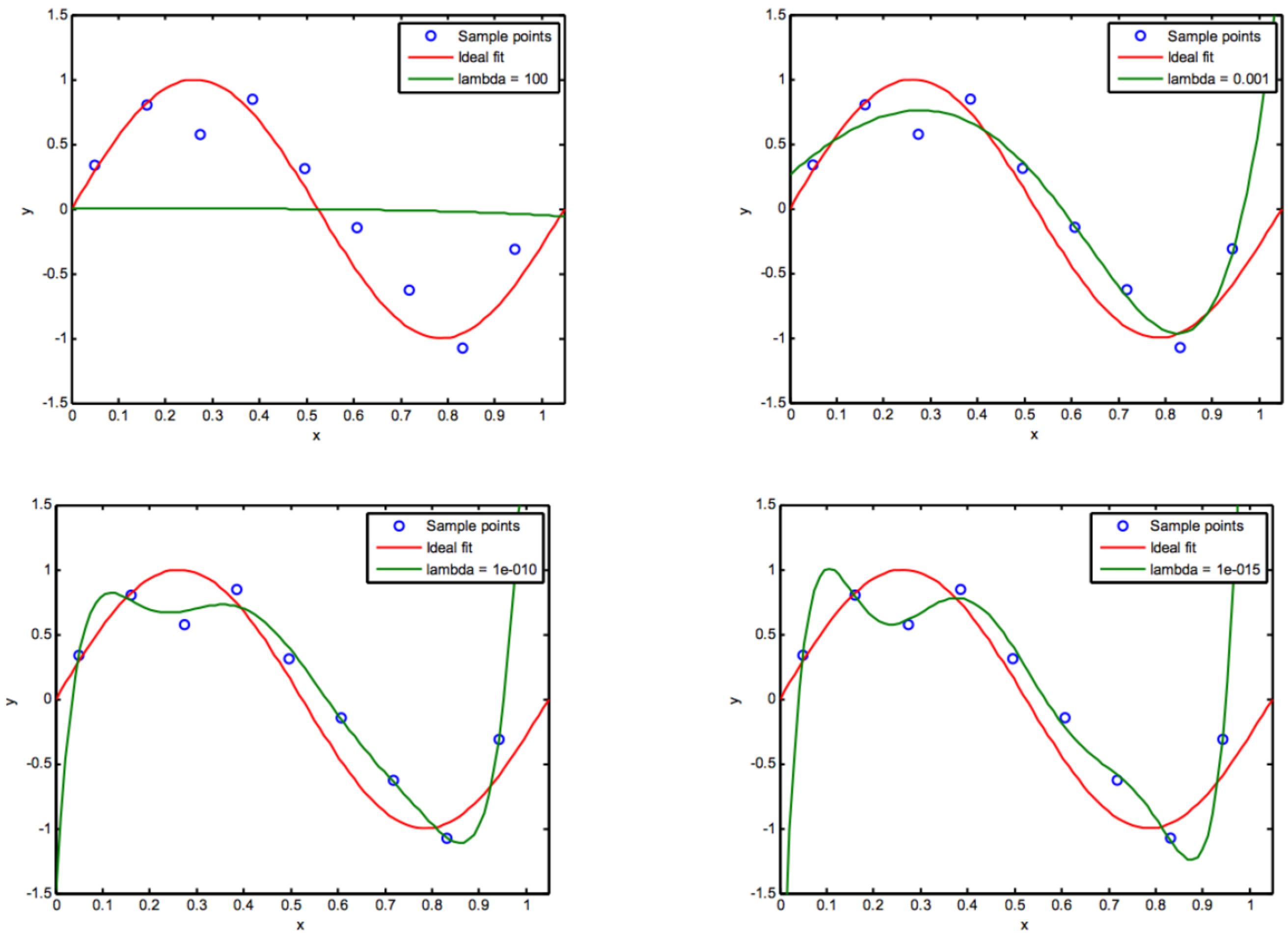
- The red curve is the true function (which is not polynomial).
- The data points are samples from the curve with added noise in y.
- There is a choice in both the degree (D) of the basis functions used and in the strength of the regularization.



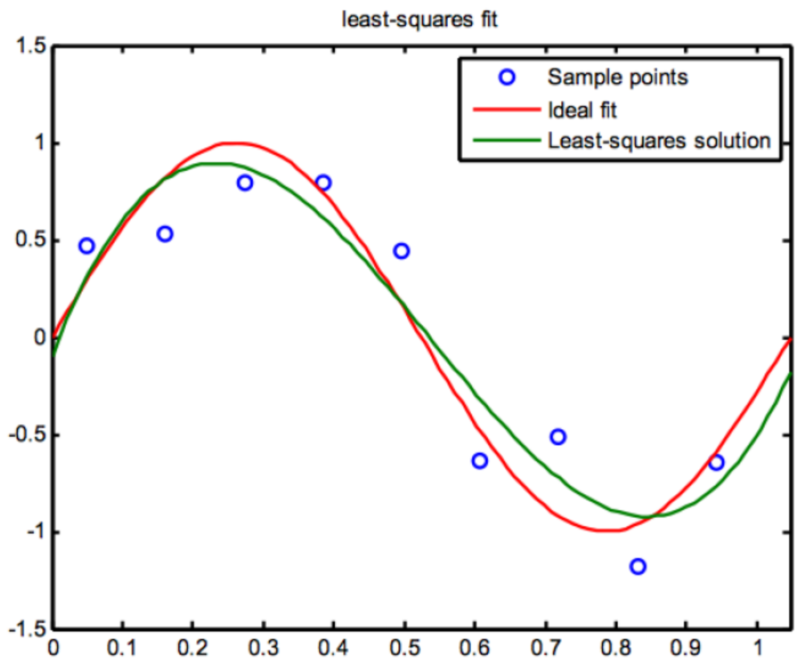
$$\tilde{E}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\{f(x_i,\theta) - y_i\}^2 + \frac{\lambda}{N}||\theta||^2$$

θ is a D+1 dimensional vector

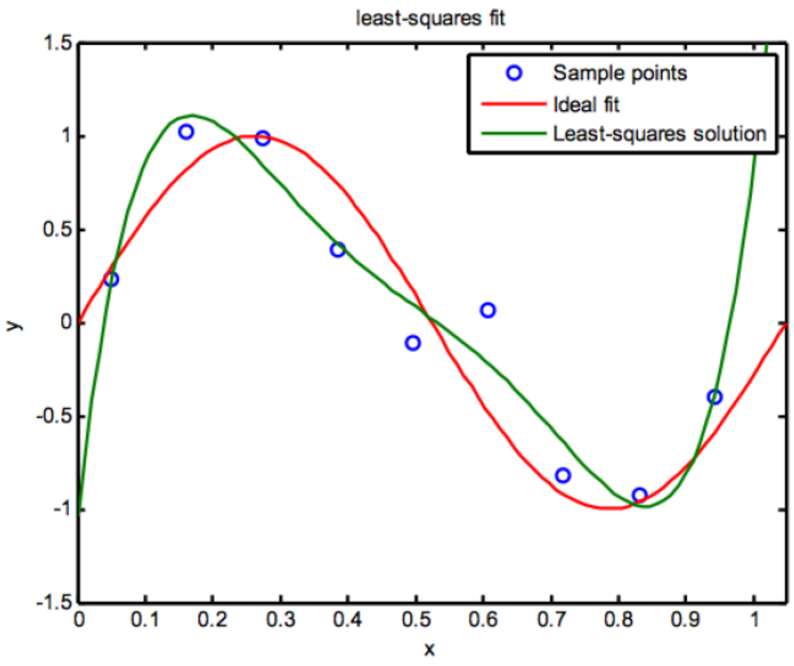
N=9 samples, D=7



N=9 samples, D=3



N=9 samples, D=5



Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization length

Regularized Regression

- Minimize with respect to

$$E(\theta) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{l(f(x_i,\theta) - y_i)}_{\text{Loss function}} + \underbrace{\lambda R(\theta)}_{\text{Regularization}}$$

- There is a choice of both loss functions and regularization.
- We have seen "ridge" regression:
  - Squared loss: $\sum_{i=1}^{N}\{f(x_i,\theta) - y_i\}^2$
  - Squared regularizer: $\lambda||\theta||^2$

The Lasso regularization (norm one)

- LASSO = Least Absolute Shrinkage and Selection

Minimize with respect to

$E(\theta) = \frac{1}{N}\sum_{i=1}^{N} l(f(x_i,\theta) - y_i) + \lambda R(\theta)$

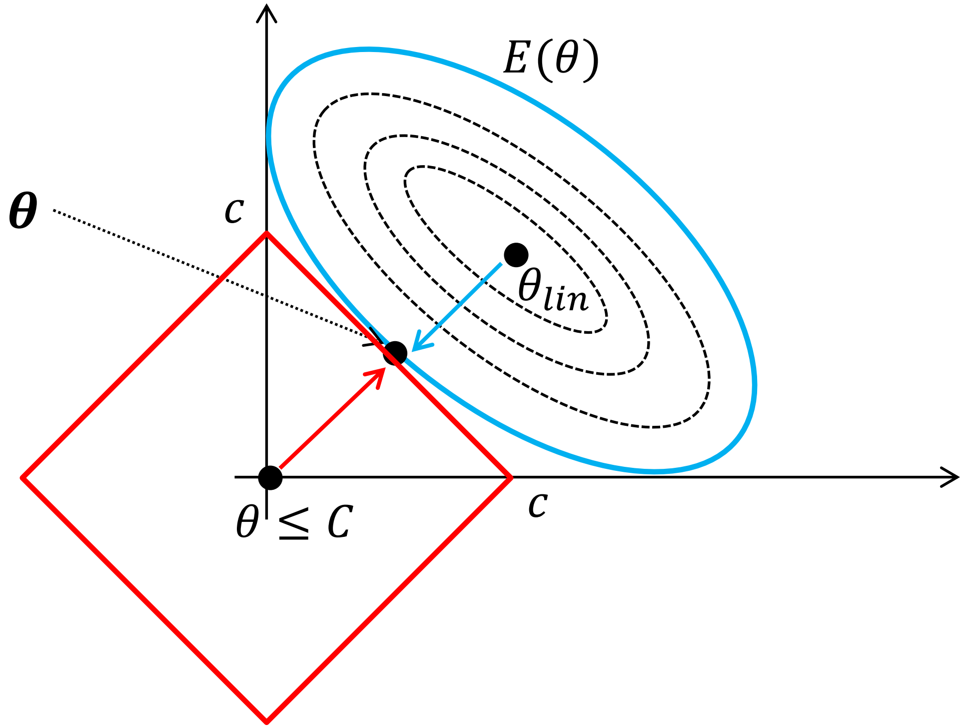
$$E(\theta) = \frac{1}{N}(y - Z\theta)^2 + \lambda||\theta||_1$$

P-Norm definition: $||\theta||_p = \left(\sum_{j=1}^{d} |\theta|^p\right)^{1/p}$

Look at an example of two parameters with Lasso

$$\begin{cases} \text{Minimize } \mathrm{E}(\theta) = \frac{1}{N}(Z\theta - y)^T(Z\theta - y) \\ \text{Subject to } \theta \leq C \end{cases}$$

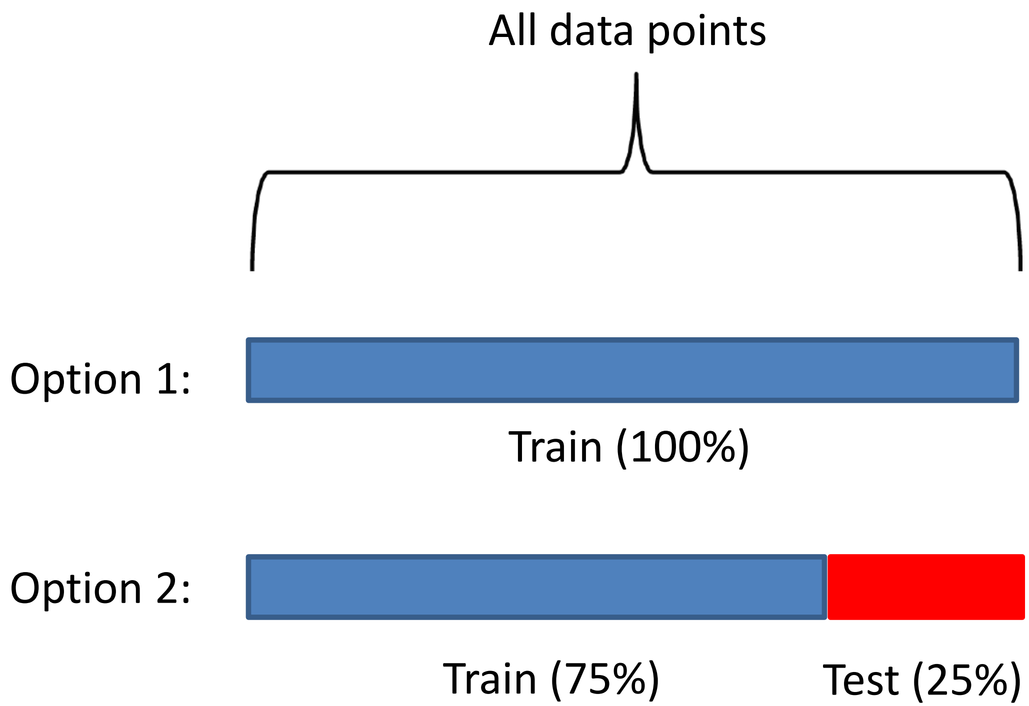
$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$



Outline

- Overfitting and regularized learning
- Ridge regression
- Lasso regression
- Determining regularization length

How to make use of data for learning?



How to do the “test”?

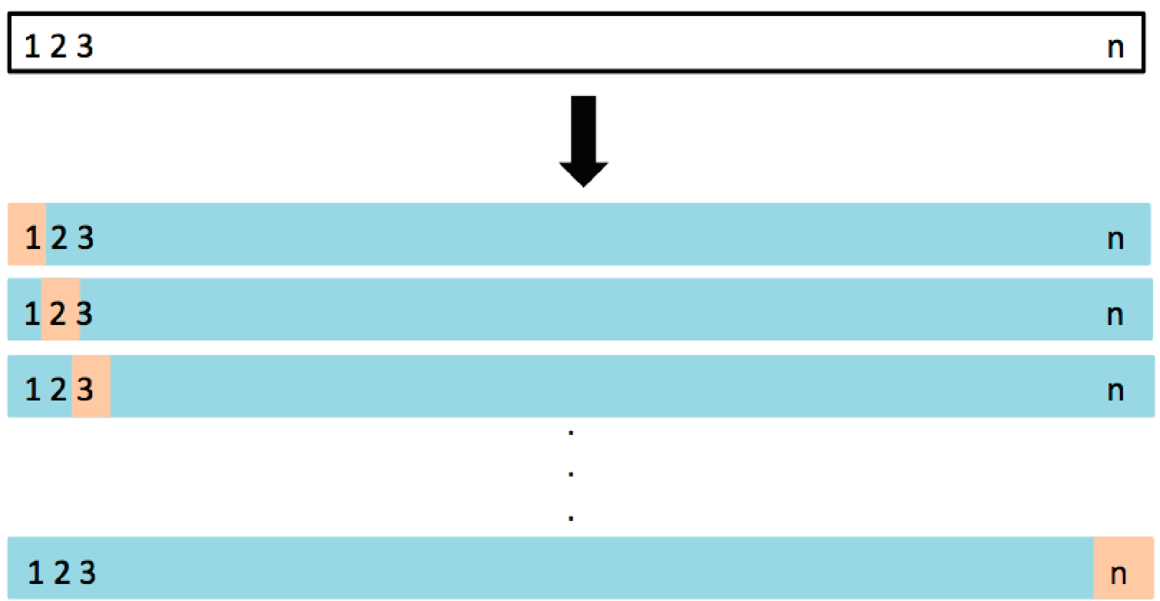
Can we have a better way?

Leave-One-Out Cross Validation

For every $i = 1, \dots, n$:

- train the model on every point except i,
- compute the test error on the held out point.

Average the test errors. $$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i^{(-i)})^2$$



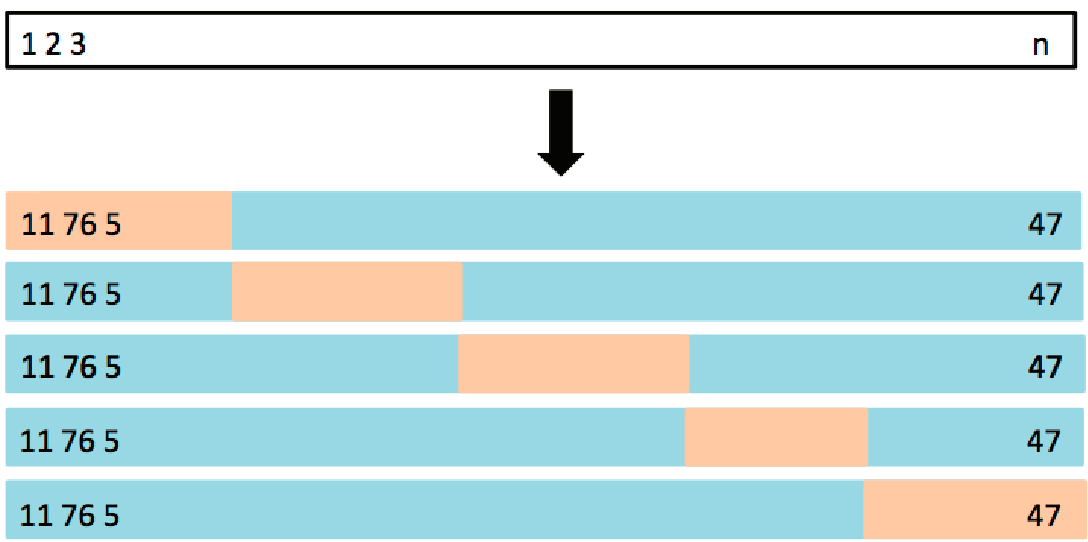
K-Fold Cross Validation

Split the data into k subsets or *folds*.

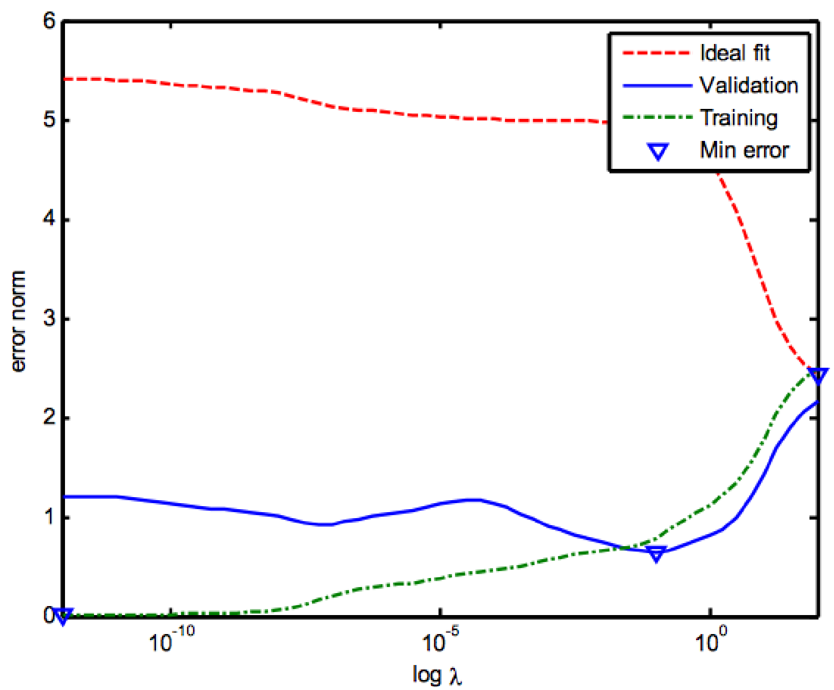
For every $i = 1, \ldots, k$:

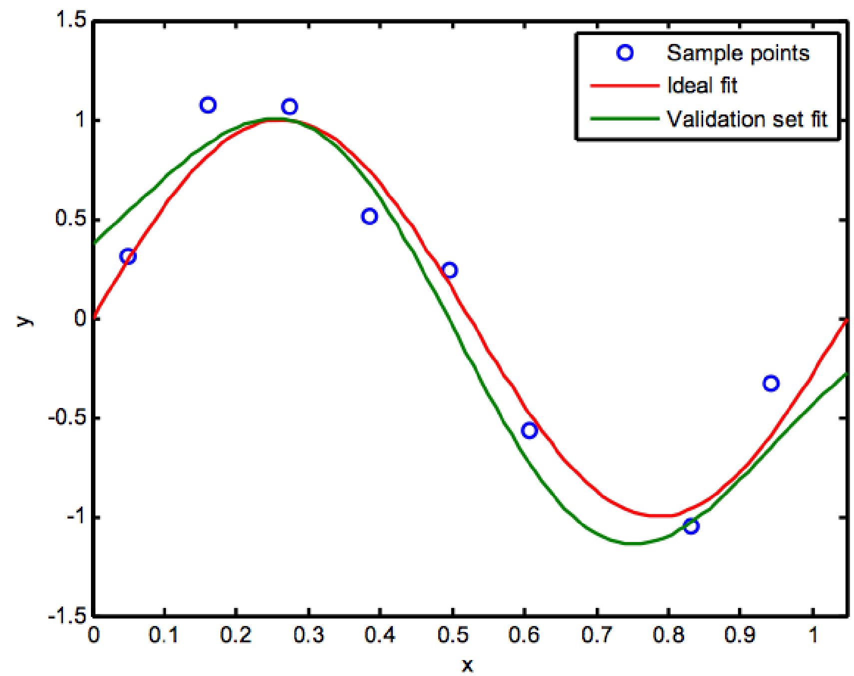
- train the model on every fold except the ith fold,
- compute the test error on the ith fold.

Average the test errors.



Choosing λ Using Validation Dataset





Pick up the lambda with the lowest mean value of RMSE calculated by Cross Validation approach

Take-Home Messages

- What is overfitting
- What is regularization
- How does Ridge regression work
- Sparsity properties of Lasso regression
- How to choose the regularization coefficient λ