18-660: Numerical Methods for Engineering Design and Optimization

Xin Li
Department of ECE
Carnegie Mellon University
Pittsburgh, PA 15213

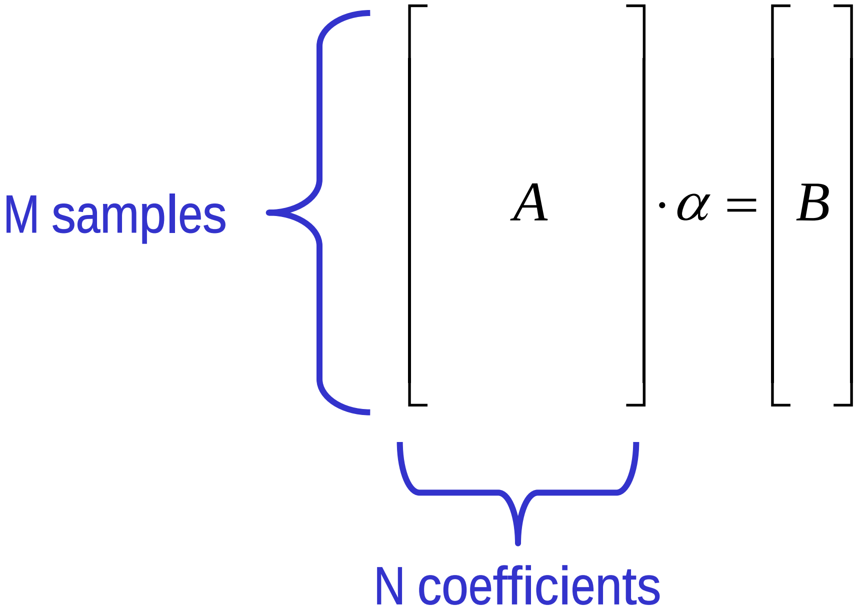
Electrical & Computer ENGINEERING

Overview

- Linear Regression
  - Over-fitting
  - Regularization
  - L_1-norm regularization

Least-Squares Regression

- Linear regression minimizes mean squared error for a set of sampling points



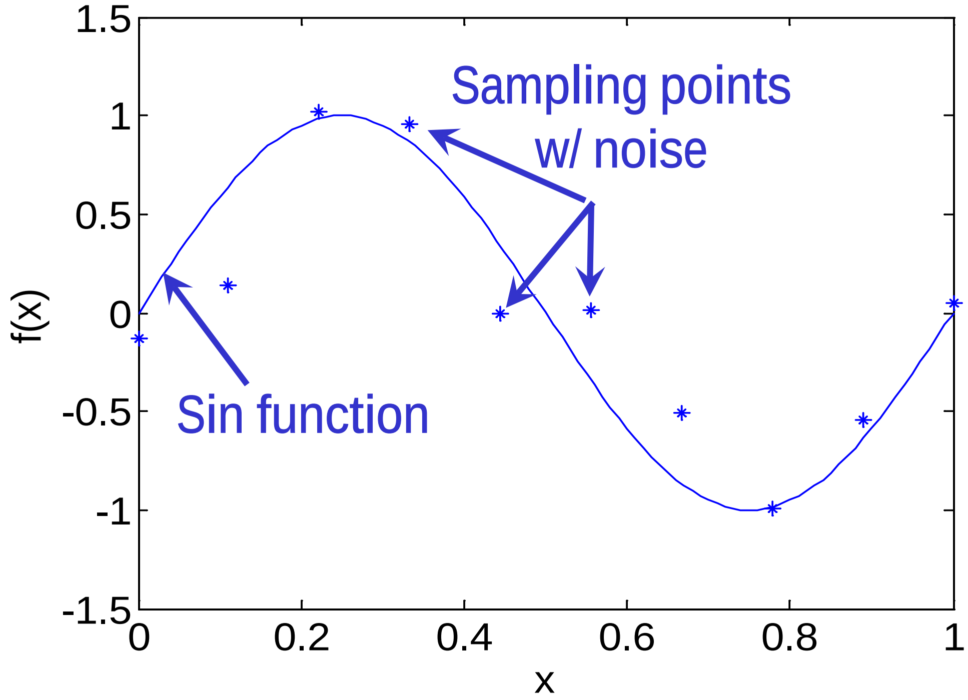
$$A \cdot \alpha = B \quad \Rightarrow \quad \min_{\alpha} \left\| A \cdot \alpha - B \right\|_2^2$$

- In practice, M must be substantially larger than N (i.e., M >> N) to avoid over-fitting

A Simple Over-Fitting Example

- Approximate sinusoidal function by polynomial model

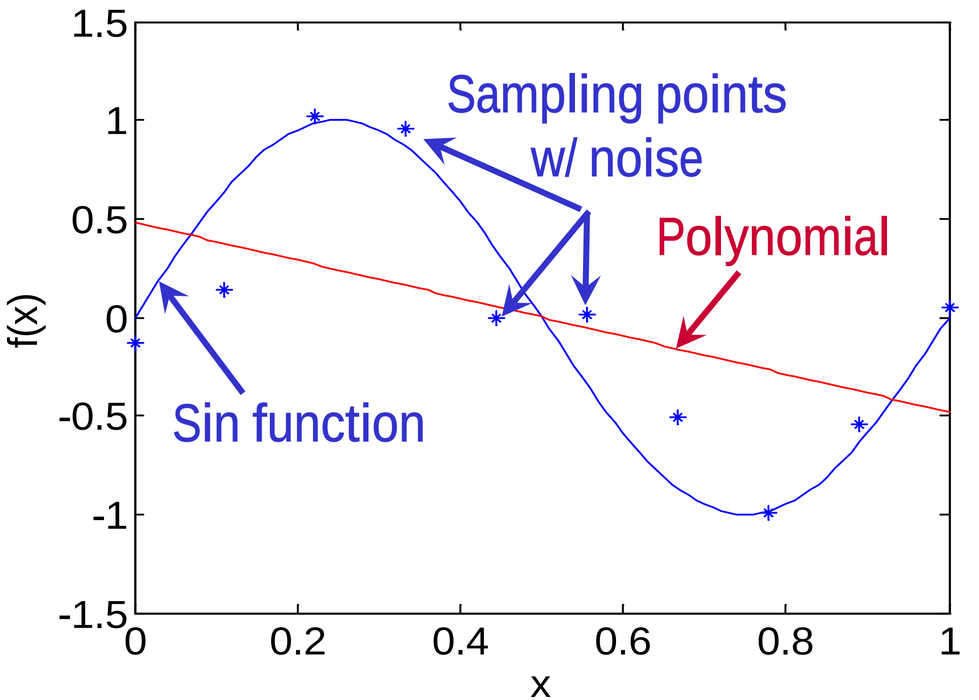
$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K} \alpha_k x^k \quad 0 \le x \le 1$$



A Simple Over-Fitting Example

- First-order polynomial model (K = 1)

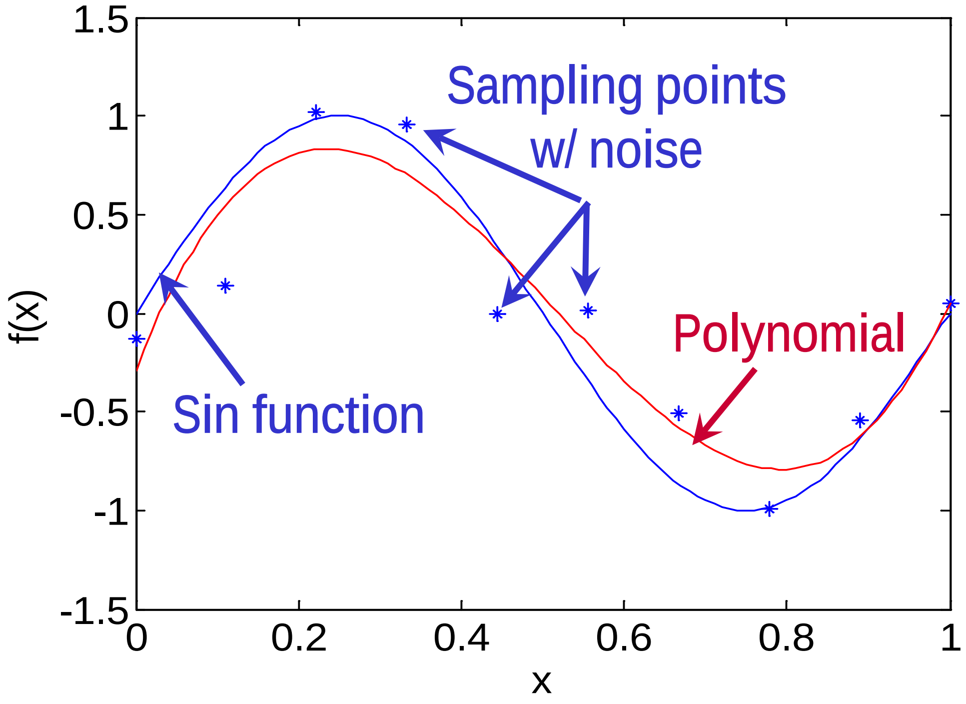
$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=1} \alpha_k x^k \quad 0 \le x \le 1$$



A Simple Over-Fitting Example

- 3rd-order polynomial model (K = 3)

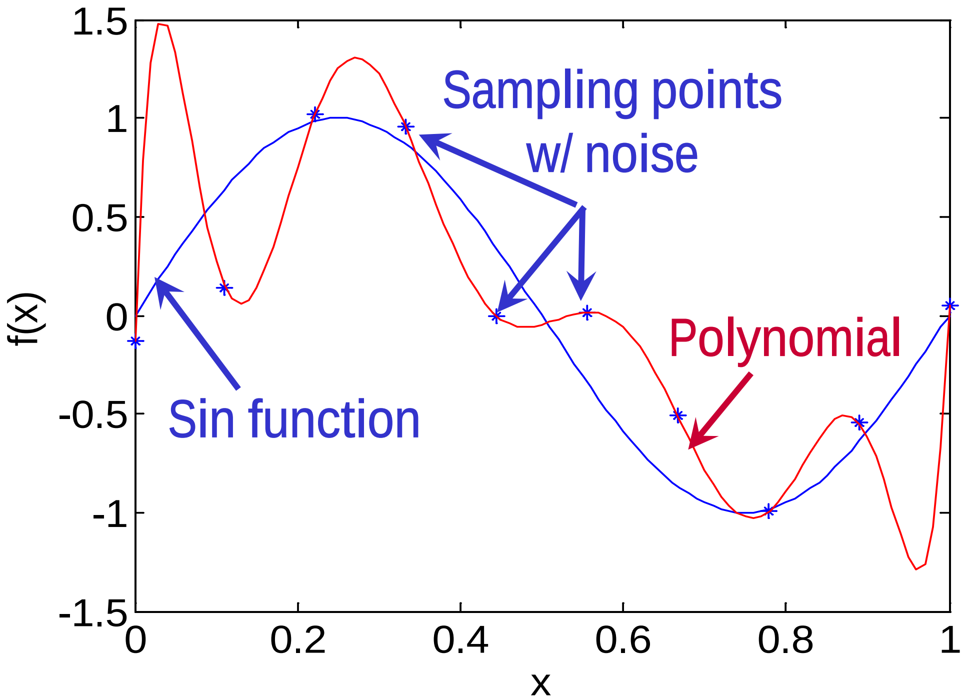
$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=3} \alpha_k x^k \quad 0 \le x \le 1$$



A Simple Over-Fitting Example

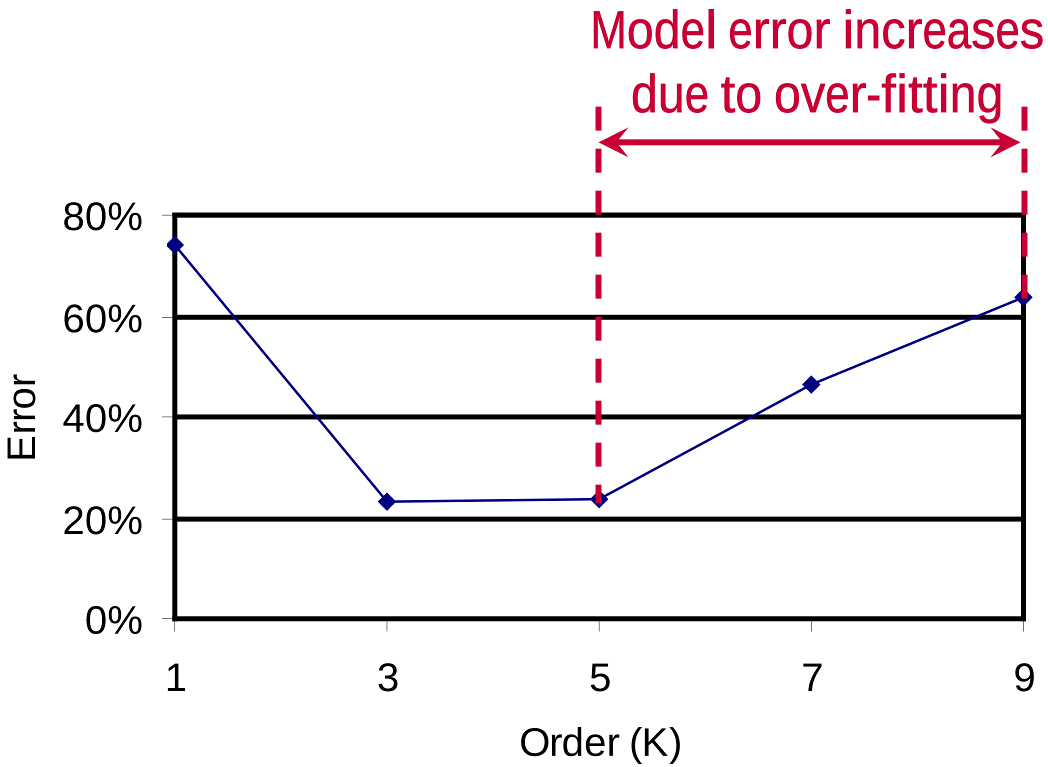
- 9rd-order polynomial model (K = 9)

$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=9} \alpha_k x^k \quad 0 \le x \le 1$$



A Simple Over-Fitting Example

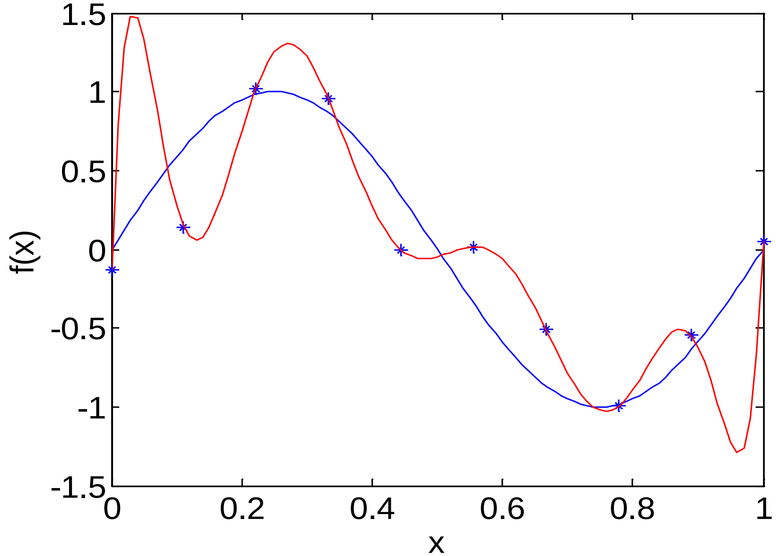
- Model order vs. model error
  - Increasing model complexity does not necessarily increase model accuracy



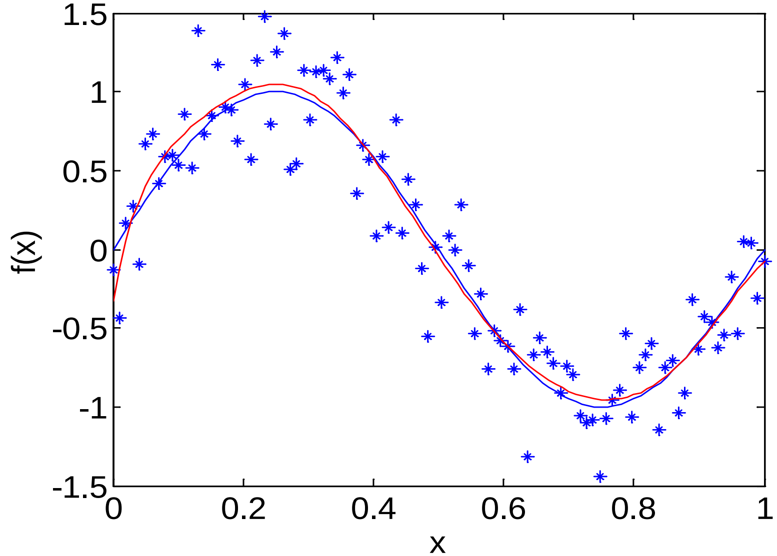
Over-Fitting

- Increasing the number of samples helps to reduce over-fitting

$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=9} \alpha_k x^k \quad 0 \le x \le 1$$



9-th order model fitted from 10 sampling points



9-th order model fitted from 100 sampling points

Regularization

- In practice, additional sampling points may be difficult and/or expensive to collect
  - E.g., a single sampling point may be collected by a physical experiment that takes several months to finish

- Regularization is a useful technique to minimize over-fitting
  - What is regularization?
  - How does it work?

Regularization

- Our previous example:

$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K} \alpha_k x^k \quad 0 \le x \le 1$$

	K = 1	K = 3	K = 9
α0	0.48	-0.29	-0.13
α1	-0.96	10.58	116.89
α2		-29.08	-2796.20
α3		18.84	26454.00
α4			-127610.00
α5			350590.00
α6			-572280.00
α7			549450.00
α8			-286530.00
α9			62610.00

Coefficients become extremely large due to over-fitting

Regularization

- Key idea: large coefficient value should be penalized

M samples

$$\begin{bmatrix} & & \\ & A & \\ & & \end{bmatrix} \cdot \alpha = \begin{bmatrix} \\ B \\ \\ \end{bmatrix}$$

N coefficients

$$\min_{\alpha} \quad \left\| A \cdot \alpha - B \right\|_2^2$$

Ordinary least-squares

$$\min_{\alpha} \quad \left\| A \cdot \alpha - B \right\|_2^2 + \lambda \cdot \left\| \alpha \right\|_2^2$$

Regularization

  - λ determines how much $||\alpha||_2$ should be penalized

Regularization

- Regularization can be re-written as a least-squares problem

$$\|A\alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_2^2 = \|A\alpha - B\|_2^2 + \lambda \cdot \alpha_1^2 + \lambda \cdot \alpha_2^2 + \cdots$$

$$\begin{aligned} A \cdot \alpha &= B \\ \sqrt{\lambda} \cdot \alpha_1 &= 0 \\ \sqrt{\lambda} \cdot \alpha_2 &= 0 \\ &\vdots \\ \sqrt{\lambda} \cdot \alpha_N &= 0 \end{aligned}$$

$$\Rightarrow \quad \begin{bmatrix} & A & \\ \hline \sqrt{\lambda} & & \\ & \sqrt{\lambda} & \\ & & \sqrt{\lambda} \end{bmatrix} \cdot \alpha = \begin{bmatrix} B \\ \hline 0 \\ 0 \\ 0 \end{bmatrix}$$

Solve least-squares solution α

Regularization

$$\min_{\alpha} \quad \left\| A \cdot \alpha - B \right\|_2^2 + \lambda \cdot \left\| \alpha \right\|_2^2$$

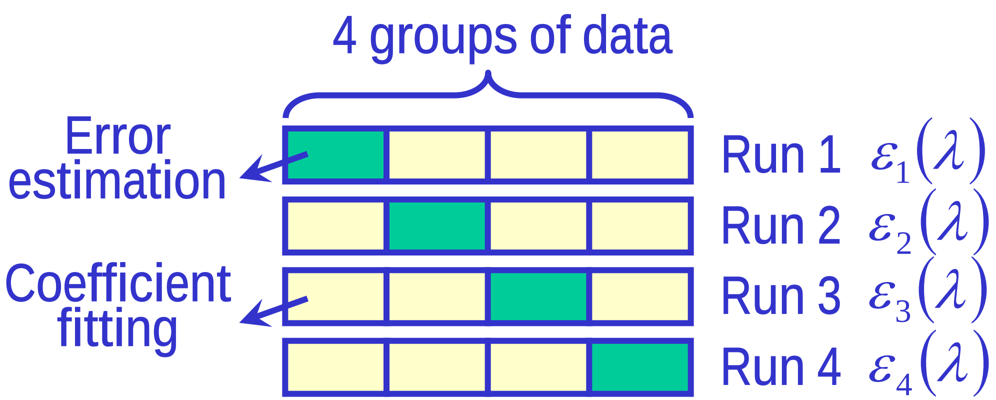
- Regularization result depends on λ
  - $\lambda = 0 \rightarrow$ ordinary least-squares regression (over-fitting)
  - $\lambda = \text{inf} \rightarrow \alpha = 0$ (over-penalized)

- Optimal λ value is case-dependent
  - Require a smart algorithm to automatically determine λ to achieve minimal modeling error
  - Question: how to estimate modeling error by using a set of training samples?

Cross Validation

- λ is often determined by cross validation
  - Calculate coefficients from training set
  - Estimate error from testing set

- Example: 4-fold cross validation



$$Error(\lambda) = [\varepsilon_1(\lambda) + \varepsilon_2(\lambda) + \varepsilon_3(\lambda) + \varepsilon_4(\lambda)]/4$$

Cross Validation

- Example: 4-fold cross validation
  - Solve the regularization problem with different λ values
  - Estimate error Error(λ) for each λ
  - Find the optimal λ to minimize Error(λ)

$$\min_{\alpha} \quad \left\| A \cdot \alpha - B \right\|_2^2 + \lambda \cdot \left\| \alpha \right\|_2^2$$

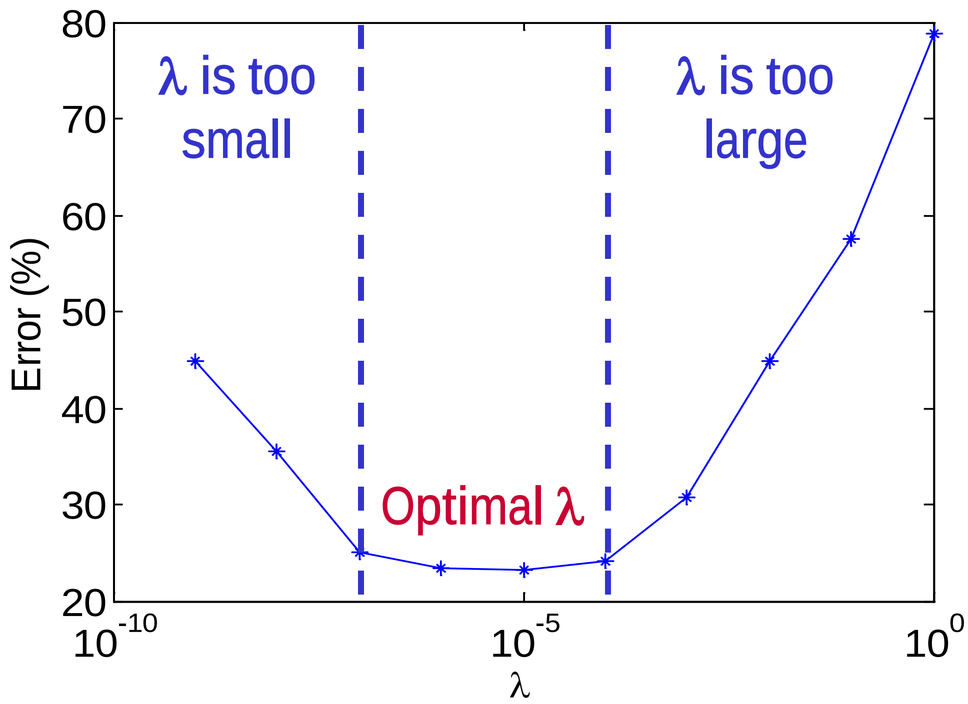
Estimate error

$$\min_{\lambda} \quad Error(\lambda) = \left[\varepsilon_1(\lambda) + \varepsilon_2(\lambda) + \varepsilon_3(\lambda) + \varepsilon_4(\lambda) \right] / 4$$

A Simple Regularization Example

- 9-th order polynomial model fitted from 10 sampling points

$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=9} \alpha_k x^k \quad 0 \le x \le 1 \qquad \min_{\alpha} \quad \|A \cdot \alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_2^2$$

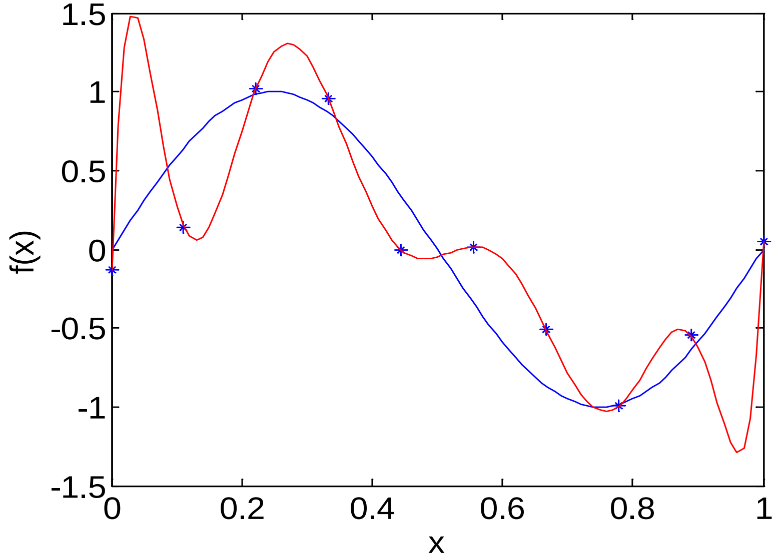


λ = 0 → ordinary least-squares
λ = inf → α = 0

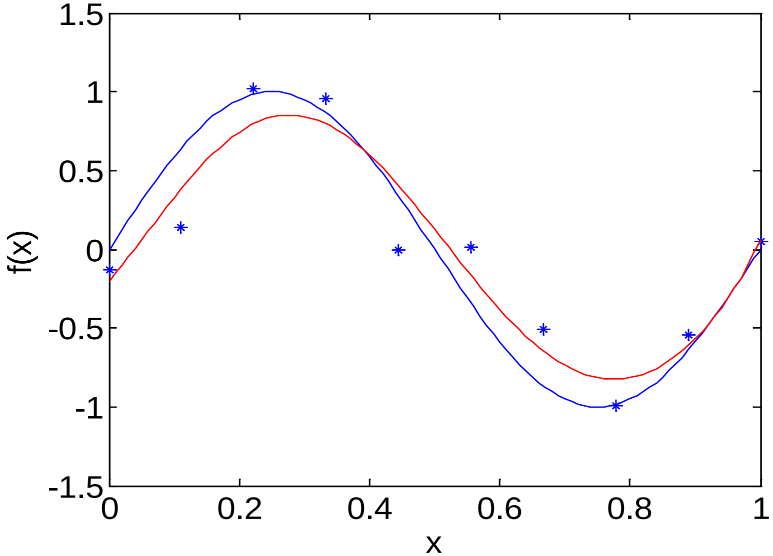
A Simple Regularization Example

- 9-th order polynomial model fitted from 10 sampling points

$$f(x) = \sin(2\pi \cdot x) \approx \sum_{k=0}^{K=9} \alpha_k x^k \quad 0 \le x \le 1 \qquad \min_{\alpha} \quad \|A \cdot \alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_2^2$$



Without regularization



With regularization (λ = 10^{-5})

Regularization

- Several other possible forms of regularization

$$\min_{\alpha} \quad \|A \cdot \alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_2^2$$

$$\min_{\alpha} \quad \|A \cdot \alpha - B\|_2^2 + \lambda \cdot \|\alpha\|_1^2$$

$$\begin{aligned} &\min_{\alpha} \quad \|A\alpha - B\|_2^2 \\ &\text{S.T.} \quad \|\alpha\|_1 \le \lambda \end{aligned}$$

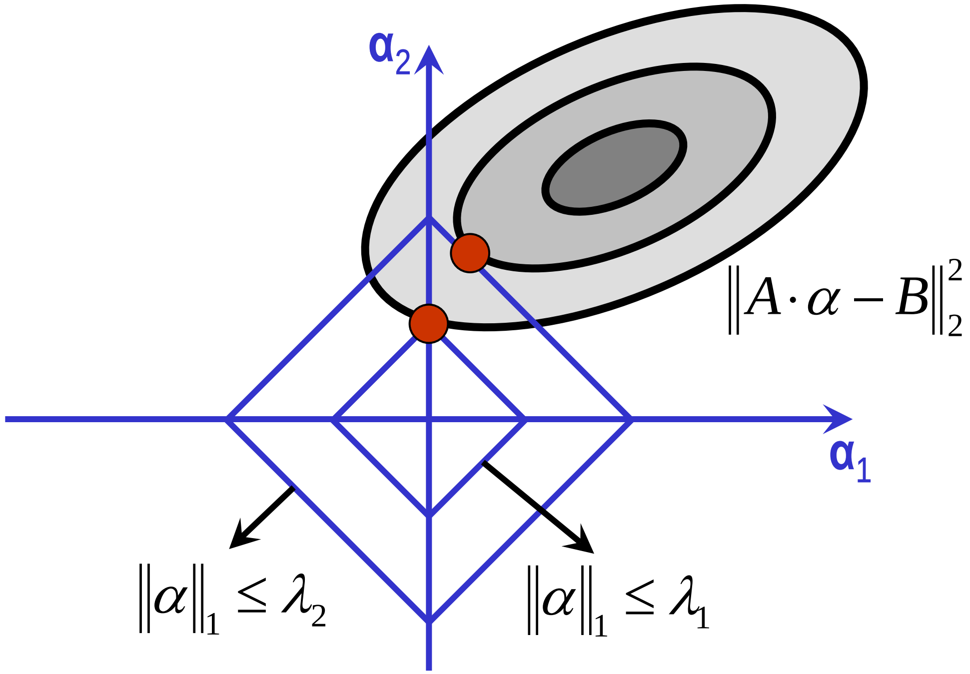
  - For a vector $\alpha \in R^N$, $||\alpha||_1$ is defined as:

$$\|\alpha\|_1 = \sum_{i=1}^{N} |\alpha_i|$$

L_1-Norm Regularization

- L_1-norm regularization is often used to find sparse solution of a linear equation

$$\min_{\alpha} \quad \|A\alpha - B\|_2^2$$
$$\text{S.T.} \quad \|\alpha\|_1 \le \lambda$$



L_1-norm regularization yields sparse solution if λ is sufficiently small

L_1-Norm Regularization

- L_1-norm regularization can be solved by convex programming

$$\min_{\alpha} \quad \|A\alpha - B\|_2^2 \qquad \text{S.T.} \quad \|\alpha\|_1 \le \lambda$$

$$\Rightarrow$$

$$\min_{\alpha,t} \quad \|A\alpha - B\|_2^2$$
$$\text{S.T.} \quad t_1 + \cdots t_N \le \lambda$$
$$-t_n \le \alpha_n \le t_n$$
$$(n = 1, 2, \cdots, N)$$

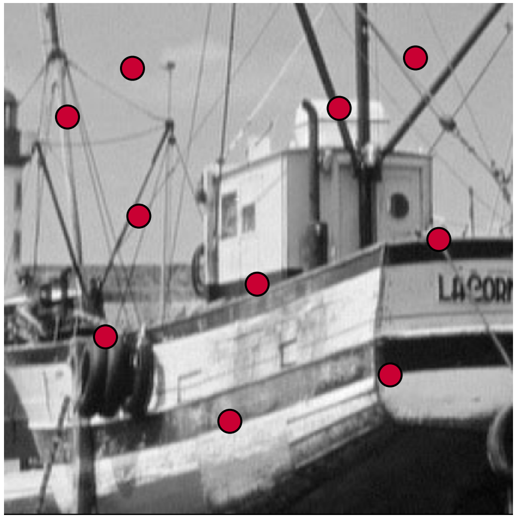
(Convex optimization)

  - Convex quadratic cost function
  - Linear constraints

L_1-norm regularization has been applied to a large number of practical problems, e.g., image processing

Image Processing Example

- Image re-construction (compressed sensing)
  - Sample an image at a small number of locations
  - Recover full image by a numerical algorithm
  - More details in next lecture…



Original image

Recovered image

Summary

- Linear regression
  - Over-fitting
  - Regularization
  - L_1-norm regularization