Linear models

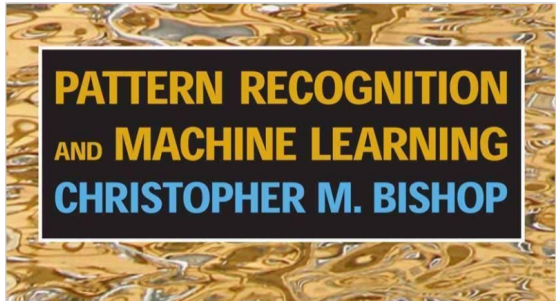
Rodrigo A. Vargas-Hernández

24/Jan/2023

References for today's lecture

(24/Jan/2023)

Chapter 3: Linear models for regression



Notation

- $\mathbf{x} \rightarrow$ single data point, $\mathbf{x} = [x_0, \cdots, x_i, \cdots, x_d]$
- $d \rightarrow$ total number of **features** in $\mathbf{x}$
- $y \rightarrow$ observable of a single point [1]
- $\mathbf{X} \rightarrow$ all data points, $\mathbf{X} = [\mathbf{x}_1^T, \cdots, \mathbf{x}_N^T]$
- $\mathbf{y} \rightarrow$ observables for all data points, $\mathbf{y} = [y_1, \cdots, y_N]$
- $N \rightarrow$ total number of data points
- $\mathcal{D} \rightarrow$ Data set, $\mathcal{D} = [\mathbf{X}, \mathbf{y}]$
- $\boldsymbol{\theta} \rightarrow$ all model's parameters
- $f(\cdot) \rightarrow$ model
- $\mathcal{L}(\cdot) \rightarrow$ loss, error or cost function

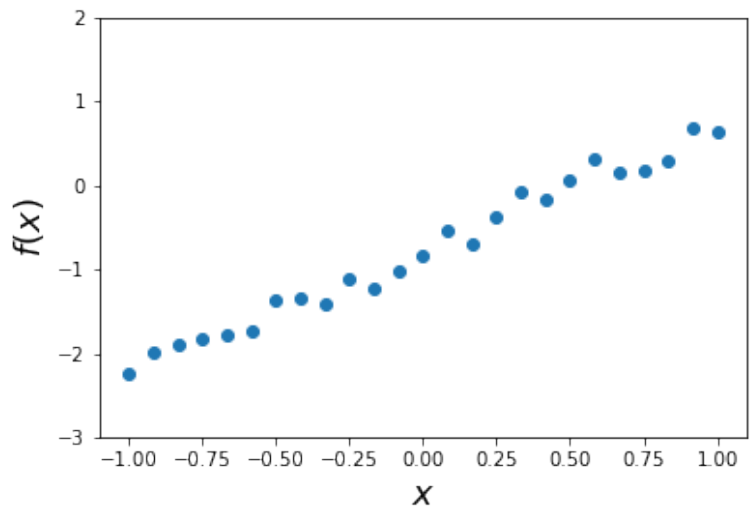
[1] Could be a vector, **multiple observables**

Vanilla model

$$f(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}}[1, \mathbf{x}] = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

- $\boldsymbol{\theta} \rightarrow$ parameters of the linear model

How can we find the optimal value of $\boldsymbol{\theta}$?

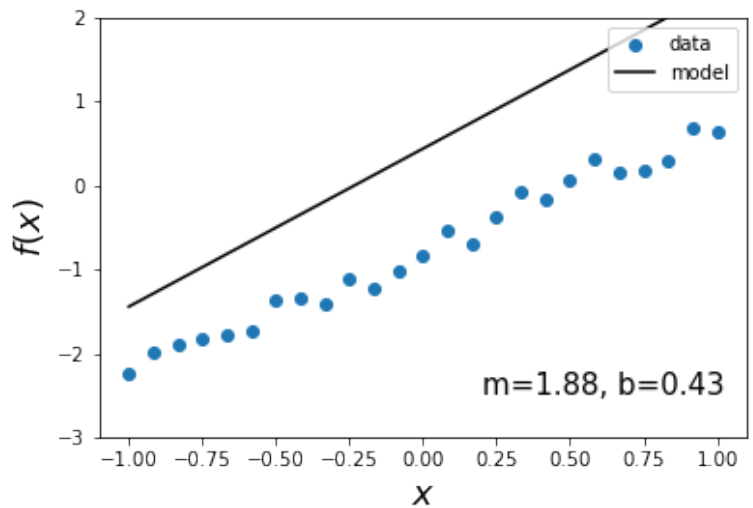


Vanilla model

$$f(\mathbf{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}}[1, \mathbf{x}] = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_d x_d = \theta_0 + \sum_{i=1}^{d} \theta_i x_i$$

- $\boldsymbol{\theta} \rightarrow$ parameters of the linear model

How can we find the optimal value of $\boldsymbol{\theta}$?



Loss function

RMSE

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i}^{N}(y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 = \frac{1}{2}\sum_{i}^{N}(y_i - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{x}_i)^2 \quad (1)$$

$$= \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \quad (2)$$

close up,

$$\mathbf{X}\boldsymbol{\theta} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} \boldsymbol{\theta} = \begin{pmatrix} [x_1^0, & \cdots, & x_1^d] \\ \cdots & \cdots & \cdots \\ [x_N^0, & \cdots, & x_N^d] \end{pmatrix} \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_d \end{pmatrix} \quad (3)$$

[1] HOMEWORK: proof and study Equation 2

Exact solution of a linear model

Gradient of a function equal to zero means a maxima or minima.

$$\nabla \mathcal{L}(\boldsymbol{\theta})\Big|_{\boldsymbol{\theta}^*} = \frac{1}{2}\nabla_{\boldsymbol{\theta}}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})\right] = 0 \tag{4}$$

To solve for $\boldsymbol{\theta}^*$, let's expand $(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})$,

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\theta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\theta}) = \mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} \tag{5}$$

$$\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left(-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}\right) = 0 \tag{6}$$

[1] HOMEWORK: proof and study Equations 5 and 6

[1] (tips): Equations from Sections 2.4.1 and 2.4.2

Exact solution of a linear model

Solving for $\boldsymbol{\theta}$,

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\left(-2\mathbf{X}^{\mathrm{T}}\mathbf{y} + 2\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta}\right) = 0 \tag{7}$$

$$\mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{8}$$

$$\boldsymbol{\theta}^{*} = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{9}$$

What is $\mathbf{X}^{\mathrm{T}}\mathbf{X}$?

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \begin{pmatrix} x_1^0 & x_2^0 & \cdots & x_N^0 \\ x_1^1 & x_2^1 & \cdots & x_N^1 \\ \vdots & \vdots & \vdots & \vdots \\ x_1^{d-1} & x_2^{d-1} & \cdots & x_N^{d-1} \\ x_1^d & x_2^d & \cdots & x_N^d \end{pmatrix} \begin{pmatrix} [x_1^0, & x_1^1, & \cdots, & x_1^{d-1}, & x_1^d] \\ [x_2^0, & x_2^1, & \cdots, & x_2^{d-1}, & x_2^d] \\ & & \cdots & & \\ [x_{N-1}^0, & x_{N-1}^1, & \cdots, & x_{N-1}^{d-1}, & x_{N-1}^d] \\ [x_N^0, & x_{N-1}^1, & \cdots, & x_N^{d-1}, & x_{Nt}^d] \end{pmatrix}$$

Optimal parameters $\rightarrow$ NumPy

$$\boldsymbol{\theta}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Operations,

- $\mathbf{X}^{\mathrm{T}} \rightarrow$ matrix transpose
- $\mathbf{X}^{\mathrm{T}}\mathbf{y} \rightarrow$ matrix-vector multiplication
- $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1} \rightarrow$ matrix inversion

Exercise

Code a general function using NumPy.

Basis-set

- What if a linear model is not enough?

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=0}^{d} \theta_i x_i$$

How many terms if we have a second-order polynomial and $d = 3$?

$$\begin{aligned}
(1 + x_1 + x_2 + x_3)^3 &= (1 + x_1 + x_2 + x_3)(1 + x_1 + x_2 + x_3)^2 \\
&= 1 + 3x_1 + 3x_1^2 + x_1^3 + 3x_2 + 6x_1x_2 + 3x_1^2x_2 \\
&\quad + 3x_2^2 + 3x_1x_2^2 + x_2^3 + 3x_3 + 6x_1x_3 + 3x_1^2x_3 \\
&\quad + 6x_2x_3 + 6x_1x_2x_3 + 3x_2^2x_3 + 3x_3^2 \\
&\quad + 3x_1x_3^2 + 3x_2x_3^2 + x_3^3
\end{aligned}$$

see this as a new representation,

$$\phi(\mathbf{x}) = [1, x_1, x_2, x_3, \cdots, x_ix_j, \cdots, x_i^m x_j^p, \cdots, x_i^m x_j^p x_\ell^r]$$

Basis-set

Linear models on basis-set expansion,

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=0}^{d} \theta_i \phi(\mathbf{x})$$

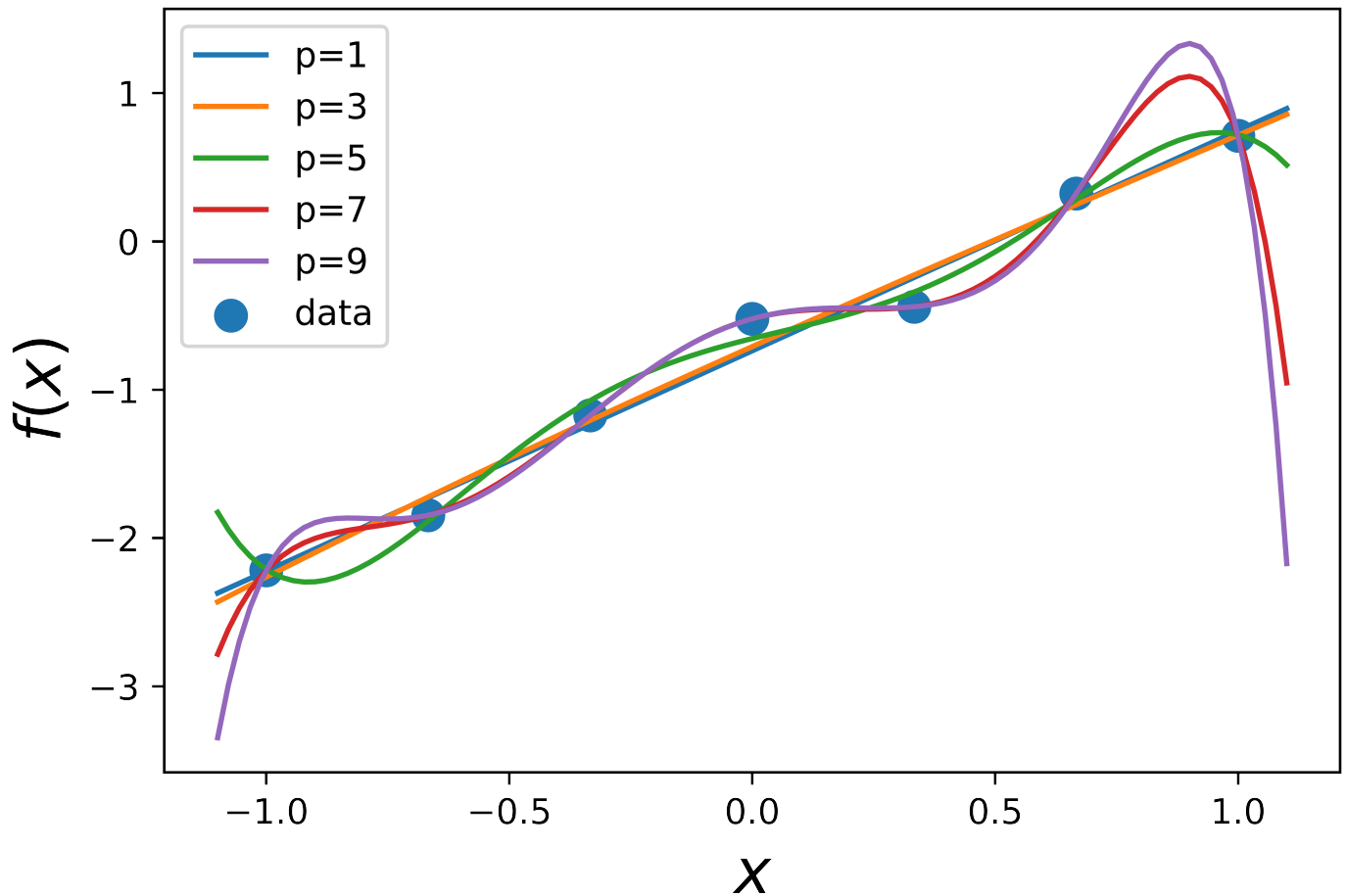
- Loss function,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}) &= \frac{1}{2} \sum_i^N (y_i - f(\mathbf{x}_i, \boldsymbol{\theta}))^2 = \frac{1}{2} \sum_i^N (y_i - \boldsymbol{\theta}^{\mathrm{T}} \phi(\mathbf{x}_i))^2 \quad (10) \\ &= \frac{1}{2} (\mathbf{y} - \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta})^{\mathrm{T}} (\mathbf{y} - \boldsymbol{\Phi}(\mathbf{X})\boldsymbol{\theta}) \quad (11) \end{aligned}$$

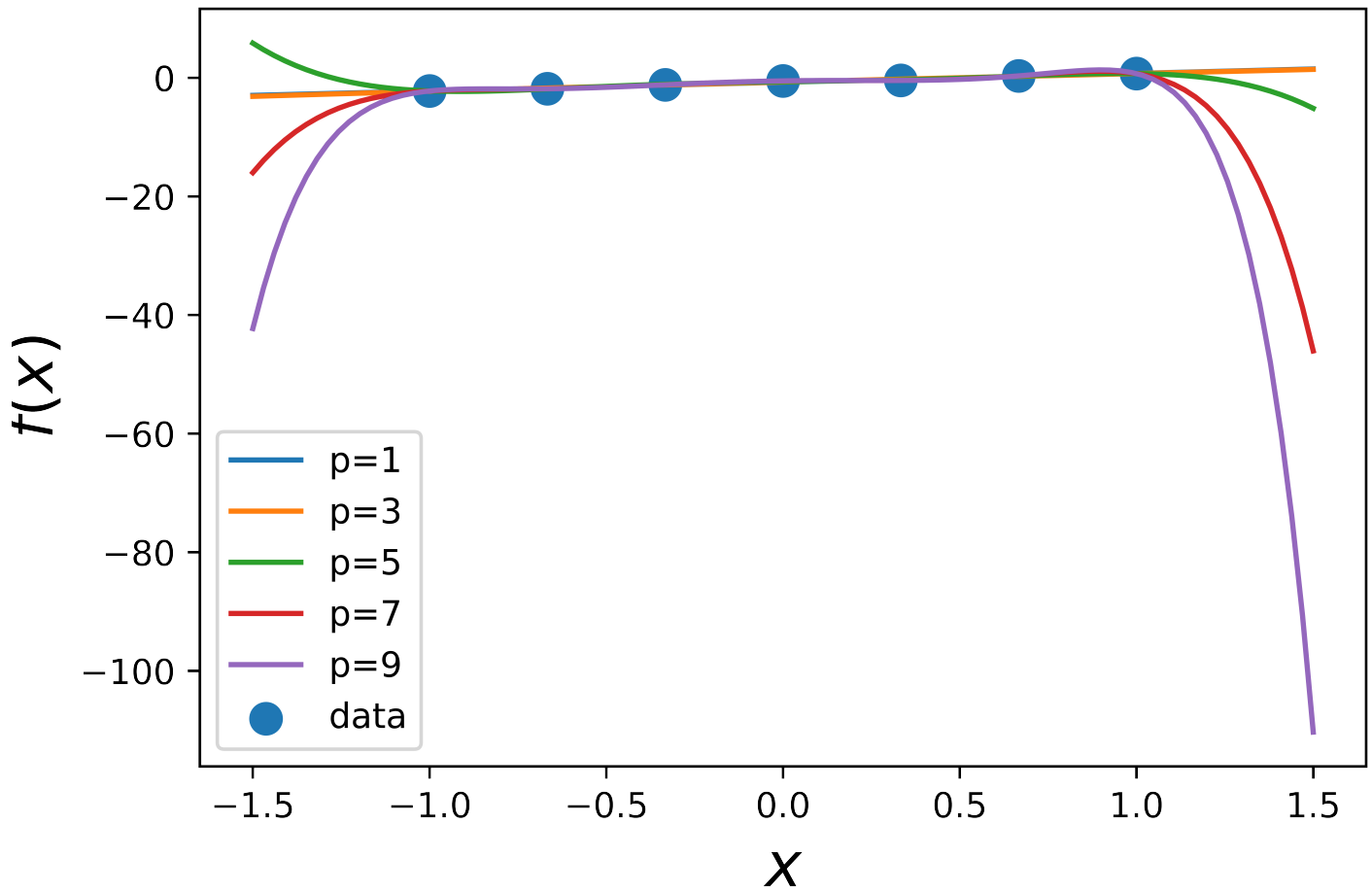
- What is $\boldsymbol{\Phi}(\mathbf{X})$?
- What is the form of the **optimal** parameters $\boldsymbol{\theta}^*$?

[1] HOMEWORK: proof and study Equations 10-11

Polynomial degree



Polynomial degree: Overfitting



Regularization Overfitting

• One technique that is often used to control the over-fitting phenomenon in such cases is that of regularization,

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i}^{N}(y_i - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_i))^2 + \frac{\lambda}{2}\sum_{j} w_j^2 \tag{12}$$

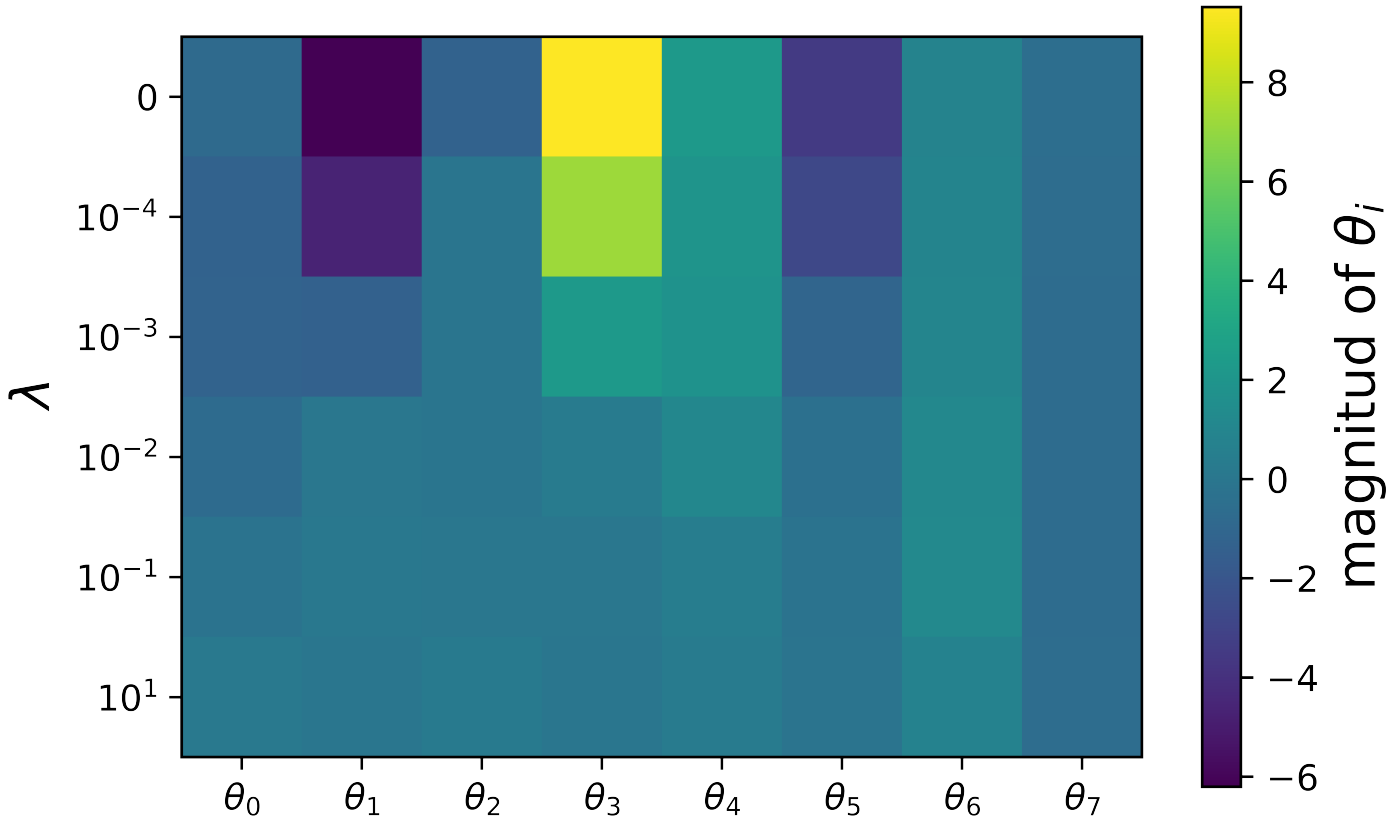
• λ is known as the **regularization** term.

(HOMEWORK)

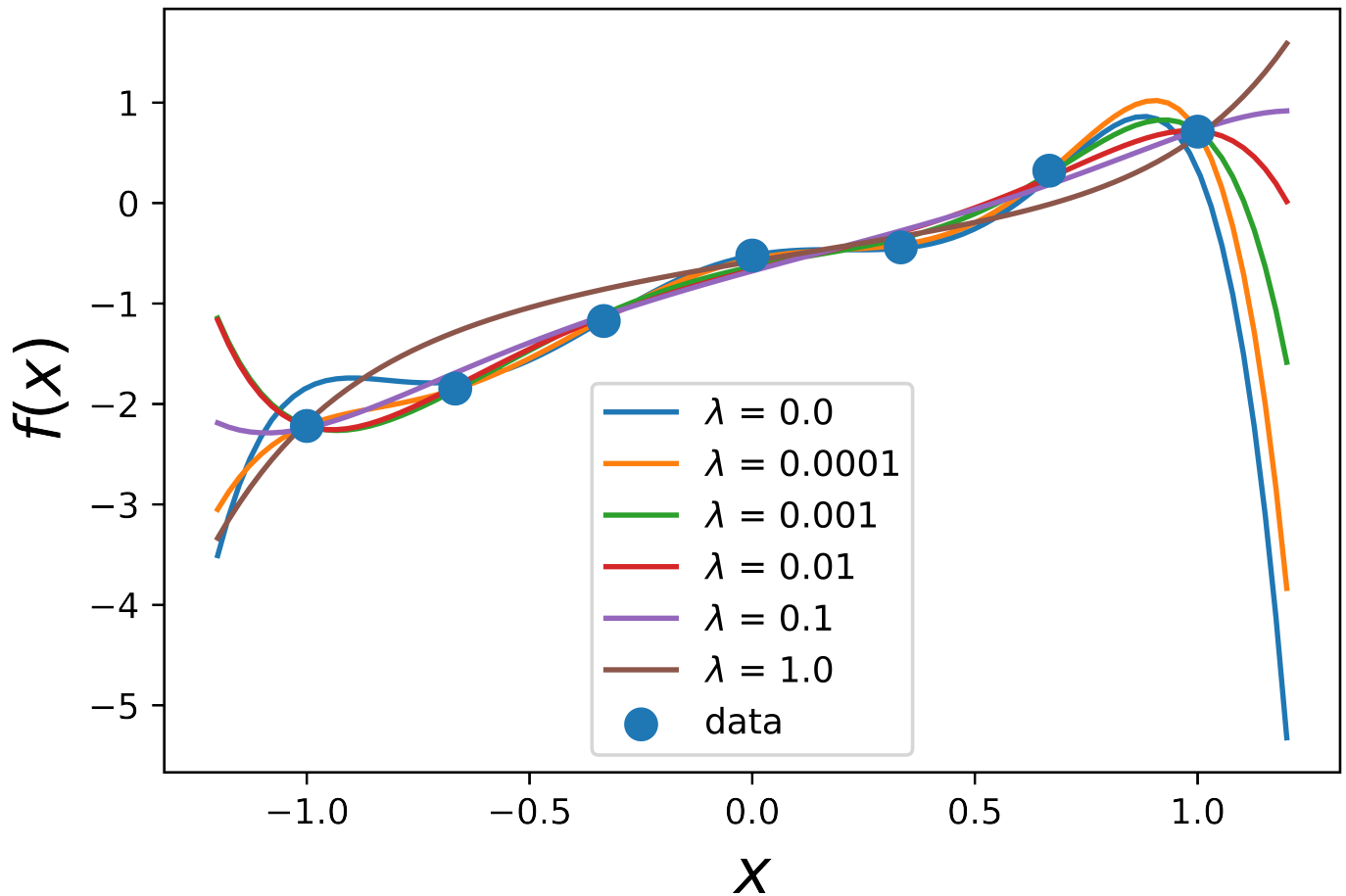
• What is the value of the **optimal** parameters $\boldsymbol{\theta}^*$ (Eq. 12)?

(tip): $\sum_j w_j^2 = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ and $\frac{\partial \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}}{\boldsymbol{w}} = 2\boldsymbol{w}$

Regularization Overfitting



Regularization Overfitting



Regularization Overfitting

Read in detail **Section 3.1.4**

- Regularization allows complex models to be trained on data sets of limited size without severe over-fitting, essentially by limiting the effective model complexity.
- The problem of determining the optimal model complexity is then shifted from one of finding the appropriate number of basis functions to one of determining a suitable value of the regularization coefficient λ.

$$\boldsymbol{\theta}^* = f(\lambda)$$