

# Lecture 4: Linear Regression (Chapter 9 of Textbook A)

Jinwoo Shin

AI503: Mathematics for AI

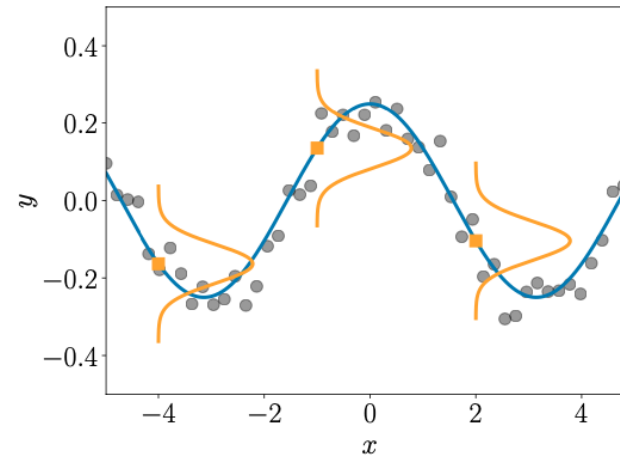
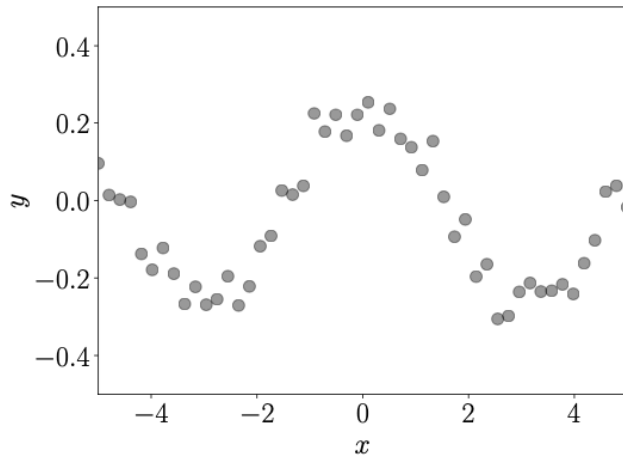
This lecture slide is based upon

<https://yung-web.github.io/home/courses/mathml.html>  
(made by Prof. Yung Yi, KAIST EE)

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression

# Regression Problem



- For some input values  $x_n$ , we observe noisy function values  $y_n = f(x_n) + \epsilon$
- Goal: infer the function  $f$  that generalizes well to function values at new inputs
- Applications: time-series analysis, control and robotics, image recognition, etc.

Notation for simplification (this is how the textbook uses)

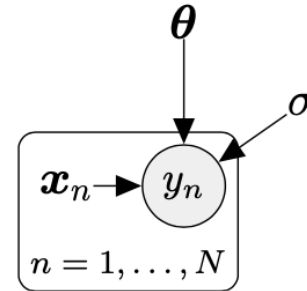
$$p(y|\mathbf{x}) = p_{Y|\mathbf{X}}(y|\mathbf{x}), \quad Y \sim \mathcal{N}(\mu, \sigma^2) \xrightarrow{\text{simplifies}} \mathcal{N}(y | f(\mathbf{x}), \sigma^2)$$

- Assume: **linear** regression, **Gaussian** noise
- $y = f(\mathbf{x}) + \epsilon$ , where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Likelihood: for  $\mathbf{x} \in \mathbb{R}^D$  and  $y \in \mathbb{R}$ ,  $p(y | \mathbf{x}) = \mathcal{N}(y | f(\mathbf{x}), \sigma^2)$
- Linear regression with the parameter  $\boldsymbol{\theta} \in \mathbb{R}^D$ , i.e.,  $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta}$

$$p(y | \mathbf{x}) = \mathcal{N}(y | \mathbf{x}^\top \boldsymbol{\theta}, \sigma^2) \iff y = \mathbf{x}^\top \boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Prior with Gaussian noise:  $p(y | \mathbf{x}) = \mathcal{N}(y | \mathbf{x}^\top \boldsymbol{\theta}, \sigma^2)$

- Training set  $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$



- Assuming iid  $N$  data samples, the likelihood is factorized into:

$$p(\mathcal{Y} | \mathcal{X}, \theta) = \prod_{n=1}^N p(y_n | \mathbf{x}_n, \theta) = \prod_{n=1}^N \mathcal{N}(y_n | \mathbf{x}_n^T, \sigma^2),$$

where  $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  and  $\mathcal{Y} = \{y_1, \dots, y_n\}$

- Estimation methods: ML and MAP

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression

# MLE (Maximum Likelihood Estimation) (1)

- $\theta_{\text{ML}} = \arg \max_{\theta} p(\mathcal{Y} | \mathcal{X}, \theta) = \arg \min_{\theta} \left( -\log p(\mathcal{Y} | \mathcal{X}, \theta) \right)$
- For Gaussian noise with  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$  and  $\mathbf{y} = [y_1, \dots, y_n]^T$ ,

$$\begin{aligned} -\log p(\mathcal{Y} | \mathcal{X}, \theta) &= -\log \prod_{n=1}^N p(y_n | \mathbf{x}_n, \theta) = -\sum_{n=1}^N \log p(y_n | \mathbf{x}_n, \theta) \\ &= \frac{1}{2\sigma^2} \sum_{n=1}^N (y_n - \mathbf{x}_n^T \theta)^2 + \text{const} = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2 + \text{const} \end{aligned}$$

Negative-log likelihood for  $f(\mathbf{x}) = \mathbf{x}^T \theta + \mathcal{N}(0, \sigma^2)$ :

$$-\log p(\mathcal{Y} | \mathcal{X}, \theta) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2 + \text{const}$$



# MLE (Maximum Likelihood Estimation) (2)

- For Gaussian noise with  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]^T$  and  $\mathbf{y} = [y_1, \dots, y_n]^T$ ,

$$\theta_{\text{ML}} = \arg \min_{\theta} \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2, \quad L(\theta) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{X}\theta\|^2$$

- In case of Gaussian noise,  $\theta_{\text{ML}} = \theta$  that minimizes the empirical risk with the squared loss function
  - Models as functions = Model as probabilistic models

# MLE (Maximum Likelihood Estimation) (3)

- We find  $\boldsymbol{\theta}$  such that  $\frac{dL}{d\boldsymbol{\theta}} = 0$

$$\frac{dL}{d\boldsymbol{\theta}} = \frac{1}{2\sigma^2} \left( -2(\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T \mathbf{X} \right) = \frac{1}{\sigma^2} \left( -\mathbf{y}^T \mathbf{X} + \boldsymbol{\theta}^T \mathbf{X}^T \mathbf{X} \right) = 0$$

$$\iff \boldsymbol{\theta}_{\text{ML}}^T \mathbf{X}^T \mathbf{X} = \mathbf{y}^T \mathbf{X}$$

$$\iff \boldsymbol{\theta}_{\text{ML}}^T = \mathbf{y}^T \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \quad (\mathbf{X}^T \mathbf{X} \text{ is positive definite if } \text{rk}(\mathbf{X}) = D)$$

$$\iff \boldsymbol{\theta}_{\text{ML}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

- Linear regression: Linear in terms of **the parameters**
  - $\phi(\mathbf{x})^\top \boldsymbol{\theta}$  is also fine, where  $\phi(\mathbf{x})$  can be non-linear (we will cover this later)
  - $\phi(\mathbf{x})$  are the features
- Linear regression with the parameter  $\boldsymbol{\theta} \in \mathbb{R}^K$ ,  $\phi(\mathbf{x}) : \mathbb{R}^D \mapsto \mathbb{R}^K$ :

$$p(y | \mathbf{x}) = \mathcal{N}(y | \phi(\mathbf{x})^\top \boldsymbol{\theta}, \sigma^2) \iff y = \phi(\mathbf{x})^\top \boldsymbol{\theta} + \epsilon = \sum_{k=0}^{K-1} \theta_k \phi_k(\mathbf{x}) + \epsilon$$

- **Example. Polynomial regression.** For  $x \in \mathbb{R}$  and  $\boldsymbol{\theta} \in \mathbb{R}^K$ , we lift the original 1-D input into  $K$ -D feature space with monomials  $x^k$ :

$$\phi(x) = \begin{pmatrix} \phi_0(x) \\ \vdots \\ \phi_{K-1}(x) \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ x^{K-1} \end{pmatrix} \in \mathbb{R}^K \implies f(x) = \sum_{k=0}^{K-1} \theta_k x^k$$

# Feature Matrix and MLE

- Now, for the entire training set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ ,

$$\mathbf{\Phi} := \begin{pmatrix} \phi^\top(\mathbf{x}_1) \\ \vdots \\ \phi^\top(\mathbf{x}_N) \end{pmatrix} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \cdots & \phi_{K-1}(\mathbf{x}_1) \\ \vdots & \cdots & \vdots \\ \phi_0(\mathbf{x}_N) & \cdots & \phi_{K-1}(\mathbf{x}_N) \end{pmatrix} \in \mathbb{R}^{N \times K}, \quad \Phi_{ij} = \phi_j(\mathbf{x}_i), \quad \phi_j : \mathbb{R}^D \mapsto \mathbb{R}$$

- Negative log-likelihood: Similarly to the case of  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$ ,

- $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y} | \mathbf{\Phi}\boldsymbol{\theta}, \sigma^2 \mathbf{I})$

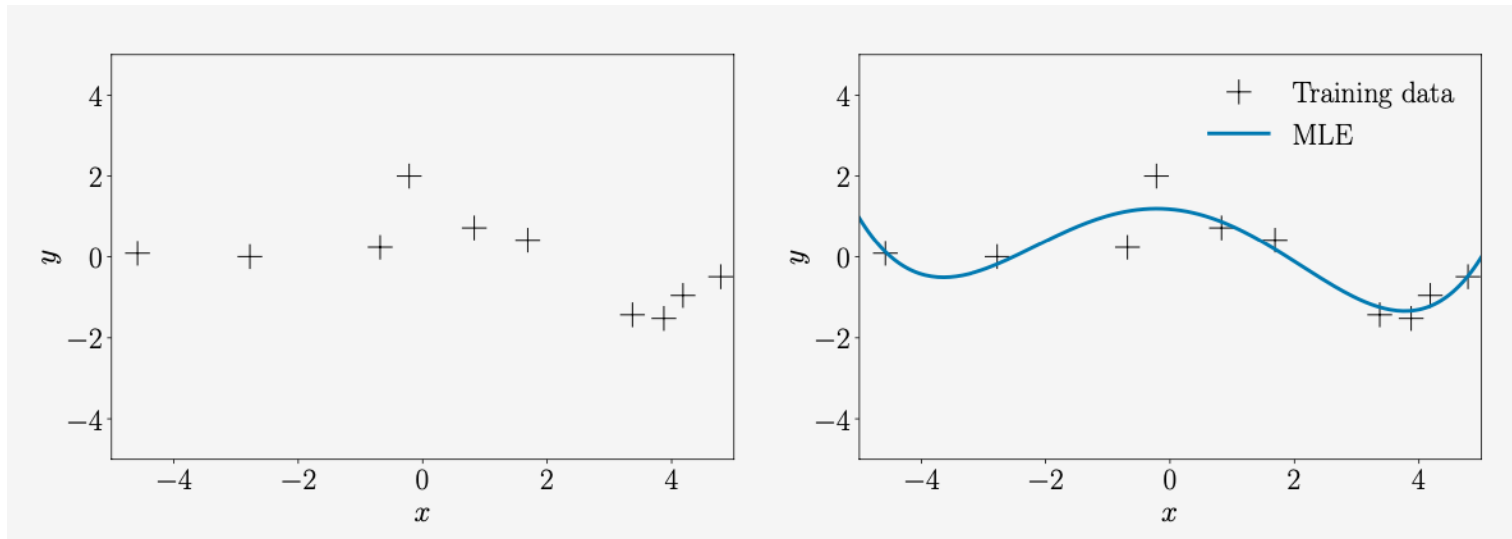
- Negative-log likelihood for  $f(\mathbf{x}) = \phi^\top(\mathbf{x})\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$ :

$$-\log p(\mathcal{Y} | \mathcal{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \mathbf{\Phi}\boldsymbol{\theta}\|^2 + \text{const}$$

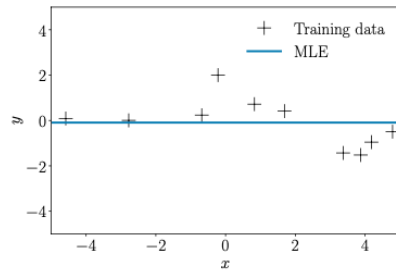
- MLE:  $\boldsymbol{\theta}_{\text{ML}} = (\mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{y}$

# Polynomial Fit

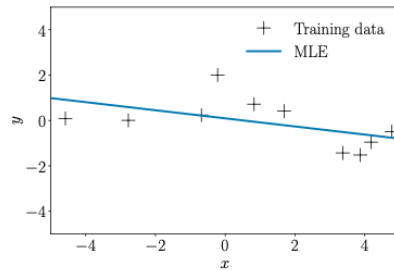
- $N = 10$  data, where  $x_n \sim \mathcal{U}[-5, 5]$  and  $y_n = -\sin(x_n/5) + \cos(x_n) + \epsilon$ ,  $\epsilon \sim \mathcal{N}(0, 0.2^2)$
- Fit with polynomial with degree 4 using ML



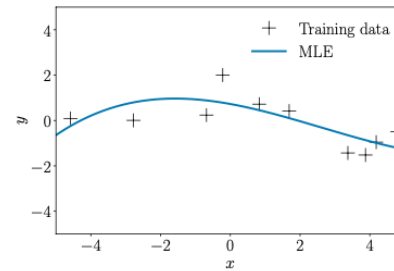
# Overfitting in Linear Regression



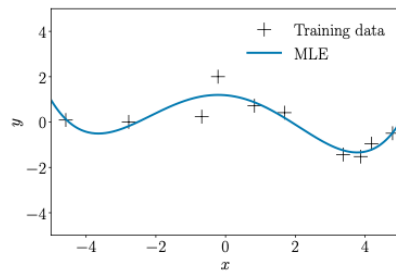
(a)  $M = 0$



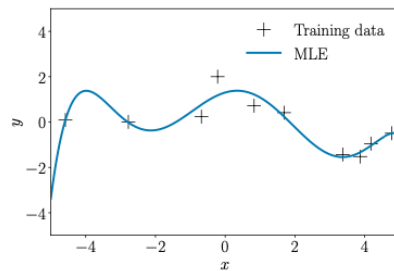
(b)  $M = 1$



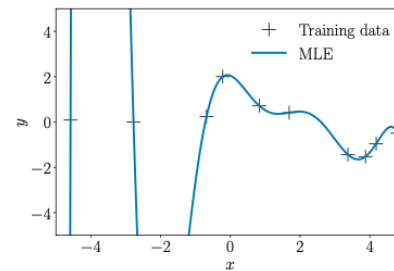
(c)  $M = 3$



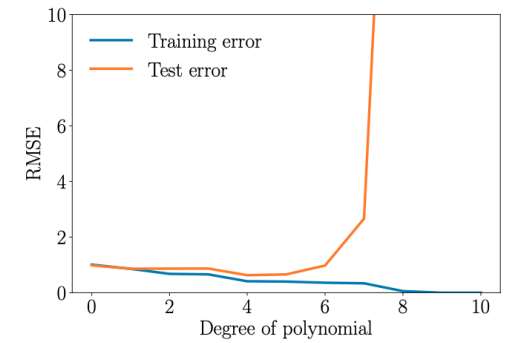
(d)  $M = 4$



(e)  $M = 6$



(f)  $M = 9$



- Higher polynomial degree is better (training error always decreases)
- Test error increases after some polynomial degree

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) **Parameter Estimation: MAP**
- (4) Bayesian Linear Regression

# MAPE (Maximum A Posteriori Estimation)

- MLE: prone to overfitting, where the magnitude of the parameters becomes large.
- a prior distribution  $p(\theta)$  helps: what  $\theta$  is plausible
- MAPE and Bayes' theorem

$$p(\theta | \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} | \mathcal{X}, \theta)p(\theta)}{p(\mathcal{Y} | \mathcal{X})} \implies \theta_{\text{MAP}} \in \arg \min_{\theta} \left( -\log p(\mathcal{Y} | \mathcal{X}, \theta) - \log p(\theta) \right)$$

- Gradient

$$-\frac{d \log p(\theta | \mathcal{X}, \mathcal{Y})}{d\theta} = -\frac{d \log p(\mathcal{Y} | \mathcal{X}, \theta)}{d\theta} - \frac{d \log p(\theta)}{d\theta}$$



# MAPE for Gaussian Prior (1)

- **Example.** A (conjugate) Gaussian prior  $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, b^2 \mathbf{I})$ 
  - For Gaussian likelihood, Gaussian prior  $\implies$  Gaussian posterior
- Negative log-posterior

L6(6)

Negative-log posterior for  $f(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$  and  $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, b^2 \mathbf{I})$ :

$$-\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top(\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \frac{1}{2b^2}\boldsymbol{\theta}^\top\boldsymbol{\theta} + \text{const}$$

- Gradient

$$-\frac{d \log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})}{d\boldsymbol{\theta}} = \frac{1}{\sigma^2}(\boldsymbol{\theta}^\top \boldsymbol{\Phi}^\top \boldsymbol{\Phi} - \mathbf{y}^\top \boldsymbol{\Phi}) + \frac{1}{b^2}\boldsymbol{\theta}^\top$$

L9(3)

# MAPE for Gaussian Prior (2)

- MAP vs. ML

$$\theta_{\text{MAP}} = \underbrace{\left( \Phi^T \Phi + \frac{\sigma^2}{b^2} I \right)}_{(*)}^{-1} \Phi^T \mathbf{y}, \quad \theta_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

- The term  $\frac{\sigma^2}{b^2} I$ 
  - Ensures that  $(*)$  is symmetric, strictly positive definite
  - Role of regularizer

## Aside: MAPE for General Gaussian Prior (3)

- **Example.** A (conjugate) Gaussian prior  $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$
- Negative log-posterior

Negative-log posterior for  $f(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x})\boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$  and  $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$ :

$$-\log p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2}(\mathbf{y} - \Phi\boldsymbol{\theta})^\top(\mathbf{y} - \Phi\boldsymbol{\theta}) + \frac{1}{2}(\boldsymbol{\theta} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1}(\boldsymbol{\theta} - \mathbf{m}_0) + \text{const}$$

- We will use this later for computing the parameter posterior distribution in Bayesian linear regression.

# Regularization: MAPE vs. Explicit Regularizer

- Explicit regularizer in regularized least squares (RLS)

$$\|\mathbf{y} - \Phi\boldsymbol{\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|^2$$

- MAPE wth Gaussian prior  $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, b^2\mathbf{I})$ 
  - Negative log-Gaussian prior

$$-\log p(\boldsymbol{\theta}) = \frac{1}{2b^2} \boldsymbol{\theta}^\top \boldsymbol{\theta} + \text{const}$$

- $\lambda = 1/2b^2$  is the regularization term
- Not surprising that we have

$$\boldsymbol{\theta}_{\text{RLS}} = \left( \Phi^\top \Phi + \lambda \mathbf{I} \right)^{-1} \Phi^\top \mathbf{y}$$

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression

# Bayesian Linear Regression

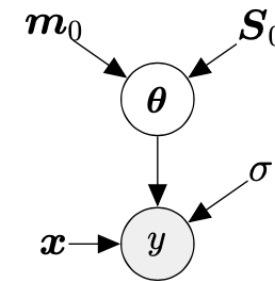
L8(4)

- Earlier, ML and MAP. Now, **fully Bayesian**
- Model

prior  $p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$

likelihood  $p(y|\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{N}(y | \phi^\top(\mathbf{x})\boldsymbol{\theta}, \sigma^2)$

joint  $p(y, \boldsymbol{\theta}|\mathbf{x}) = p(y | \mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta})$



- Goal: For an input  $\mathbf{x}_*$ , we want to compute the following **posterior predictive distribution**<sup>1</sup> of  $y_*$ :

$$p(y_*|\mathbf{x}_*, \mathcal{X}, \mathcal{Y}) = \int \underbrace{p(y_*|\mathbf{x}_*, \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y})}_{(*)} d\boldsymbol{\theta}$$

- (\*): parameter posterior distribution that needs to be computed

---

<sup>1</sup>**Chapter 9.3.4** For ease of understanding, I've slightly changed the organization of these lecture slides from that of the textbook.

# Parameter Posterior Distribution (1)

- Parameter posterior distribution

Chapter 9.3.3

$$p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) = \mathcal{N}(\boldsymbol{\theta} | \mathbf{m}_N, \mathbf{S}_N), \quad \text{where}$$
$$\mathbf{S}_N = (\mathbf{S}_0^{-1} + \sigma^2 \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \quad \mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \sigma^{-2} \boldsymbol{\Phi}^\top \mathbf{y})$$

(Proof Sketch)

- From the negative-log posterior for general Gaussian prior,

$$-\log p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) = \frac{1}{2\sigma^2} (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta})^\top (\mathbf{y} - \boldsymbol{\Phi}\boldsymbol{\theta}) + \frac{1}{2} (\boldsymbol{\theta} - \mathbf{m}_0)^\top \mathbf{S}_0^{-1} (\boldsymbol{\theta} - \mathbf{m}_0) + \text{const}$$

## Parameter Posterior Distribution (2)

$$\begin{aligned} &= \frac{1}{2} \left( \sigma^{-2} \mathbf{y}^\top \mathbf{y} - 2\sigma^{-2} \mathbf{y}^\top \Phi \boldsymbol{\theta} + \boldsymbol{\theta}^\top \sigma^{-2} \Phi^\top \Phi \boldsymbol{\theta} + \boldsymbol{\theta}^\top \mathbf{S}_0^{-1} \boldsymbol{\theta} - 2\mathbf{m}_0^\top \mathbf{S}_0^{-1} \boldsymbol{\theta} + \mathbf{m}_0^\top \mathbf{S}_0^{-1} \mathbf{m}_0 \right) \\ &= \frac{1}{2} \left( \boldsymbol{\theta}^\top (\sigma^{-2} \Phi^\top \Phi + \mathbf{S}_0^{-1}) \boldsymbol{\theta} - 2(\sigma^{-2} \Phi^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0)^\top \boldsymbol{\theta} \right) + \text{const} \end{aligned}$$

- cyan color: quadratic term, orange color: linear term
- $p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) \propto \exp(\text{quadratic in } \boldsymbol{\theta}) \implies$  Gaussian distribution
- Assume that  $p(\boldsymbol{\theta}|\mathcal{X}, \mathcal{Y}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_N, \mathbf{S}_N)$ , and find  $\mathbf{m}_N$  and  $\mathbf{S}_N$ .

$$\begin{aligned} -\log \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_N, \mathbf{S}_N) &= \frac{1}{2} (\boldsymbol{\theta} - \mathbf{m}_N)^\top \mathbf{S}_N^{-1} (\boldsymbol{\theta} - \mathbf{m}_N) + \text{const} \\ &= \frac{1}{2} \left( \boldsymbol{\theta}^\top \mathbf{S}_N^{-1} \boldsymbol{\theta} - 2\mathbf{m}_N^\top \mathbf{S}_N^{-1} \boldsymbol{\theta} + \mathbf{m}_N^\top \mathbf{S}_N^{-1} \mathbf{m}_N \right) + \text{const} \end{aligned}$$

- Thus,  $\mathbf{S}_N^{-1} = \sigma^{-2} \Phi^\top \Phi + \mathbf{S}_0^{-1}$  and  $\mathbf{m}_N^\top \mathbf{S}_N^{-1} = (\sigma^{-2} \Phi^\top \mathbf{y} + \mathbf{S}_0^{-1} \mathbf{m}_0)^\top$



- Posterior predictive distribution

L6(5)

$$\begin{aligned} p(y_* | \mathbf{x}_*, \mathcal{X}, \mathcal{Y}) &= \int p(y_* | \mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{X}, \mathcal{Y}) d\boldsymbol{\theta} \\ &= \int \mathcal{N}(y_* | \phi^\top(\mathbf{x}_*) \boldsymbol{\theta}, \sigma^2) \mathcal{N}(\boldsymbol{\theta} | \mathbf{m}_N, \mathbf{S}_N) d\boldsymbol{\theta} \\ &= \mathcal{N}(y_* | \phi^\top(\mathbf{x}_*) \mathbf{m}_N, \phi^\top(\mathbf{x}_*) \mathbf{S}_N \phi(\mathbf{x}_*) + \sigma^2) \end{aligned}$$

- The mean  $\phi^\top(\mathbf{x}_*) \mathbf{m}_N$  coincides with the MAP estimate

## Posterior Predictions (2): MAP vs. ML vs. BLR

- Assume  $p(\boldsymbol{\theta}) \sim \mathcal{N}(0, b^2 \mathbf{I})$  and  $p(y|\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{N}(y | \phi^\top(\mathbf{x})\boldsymbol{\theta}, \sigma^2)$ :

$$p(y_* | x_*, \mathcal{X}, \mathcal{Y}) = \begin{cases} \mathcal{N}(y_* | \phi^\top(\mathbf{x}_*)\boldsymbol{\theta}_{\text{ML}}, \sigma^2) & \text{for MLE} \\ \mathcal{N}(y_* | \phi^\top(\mathbf{x}_*)\boldsymbol{\theta}_{\text{MAP}}, \sigma^2) & \text{for MAP,} \\ \mathcal{N}(y_* | \phi^\top(\mathbf{x}_*)\boldsymbol{\theta}_{\text{MAP}}, \phi^\top(\mathbf{x}_*)\mathbf{S}_N\phi(\mathbf{x}_*) + \sigma^2) & \text{for BLR} \end{cases}$$

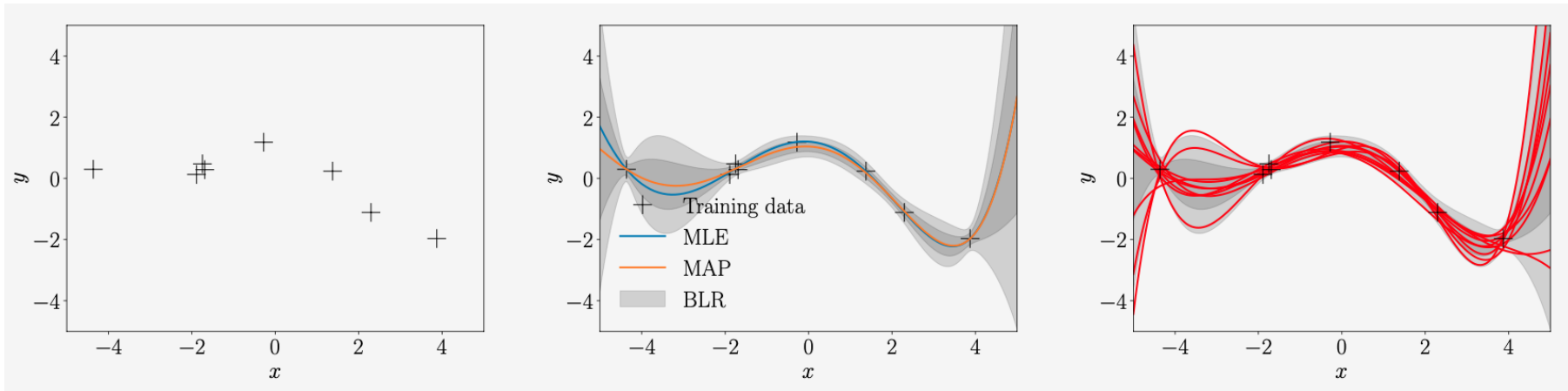
where BLR = Bayesian Linear Regression and

$$\boldsymbol{\theta}_{\text{MAP}} = \left( \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \frac{\sigma^2}{b^2} \mathbf{I} \right)^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

$$\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

$$\mathbf{S}_N = \left( \frac{1}{b^2} \mathbf{I} + \sigma^2 \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}$$

# Posterior Predictions (3)



# Computing Marginal Likelihood

- Likelihood:  $p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})$ , Marginal likelihood:  $p(\mathcal{Y}|\mathcal{X}) = \int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$

$$\begin{aligned} p(\mathcal{Y}|\mathcal{X}) &= \int p(\mathcal{Y}|\mathcal{X}, \boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta} = \int \mathcal{N}(\mathbf{y}|\Phi\boldsymbol{\theta}, \sigma^2\mathbf{I})\mathcal{N}(\boldsymbol{\theta}|\mathbf{m}_0, \mathbf{S}_0) d\boldsymbol{\theta} \\ &= \mathcal{N}(\mathbf{y} | \Phi\mathbf{m}_0, \Phi\mathbf{S}_0\Phi^T + \sigma^2\mathbf{I}) \end{aligned}$$

Questions?