

QUY NHON UNIVERSITY
DEPARTMENT OF MATHEMATICS AND STATISTICS

Linear Regression

Nguyen Quoc Duong

Master's Student, Department of Mathematics and Statistics, Quy Nhon University

April 1, 2023



Warm-Up

Please watch this tutorial video by Luis Serrano on PCA.

<https://www.youtube.com/watch?v=wYPUhge9w5c>



Roadmap

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection

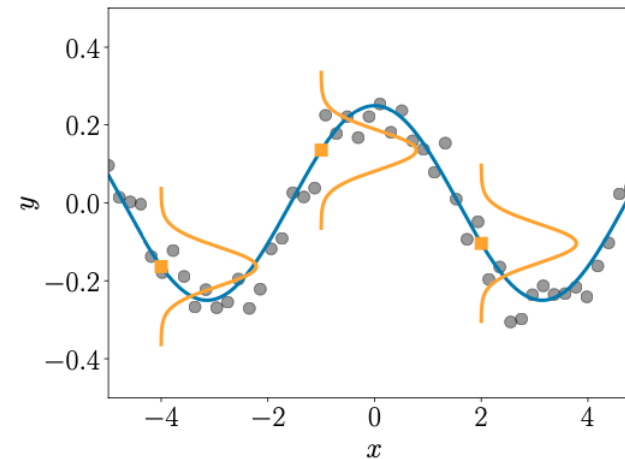
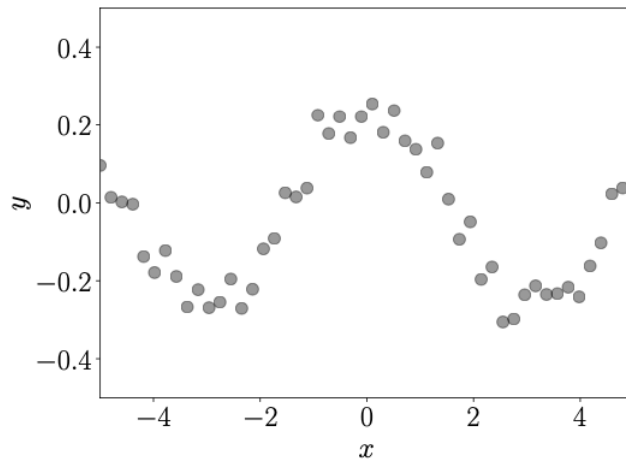


Roadmap

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection



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.



Formulation

Notation for simplification (this is how the textbook uses)

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

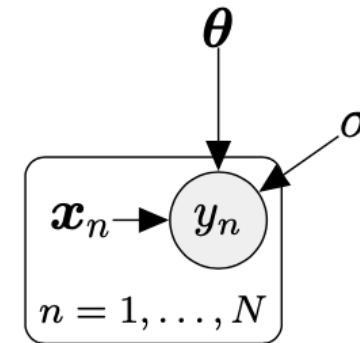
- Assume: linear regression, Gaussian noise
- $y = f(\mathbf{x}) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Likelihood: for $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$, $p(y \mid \mathbf{x}) = \mathcal{N}(y \mid f(\mathbf{x}), \sigma^2)$
- Linear regression with the parameter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

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

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



Parameter Estimation



- 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:

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

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

- Estimation methods: ML and MAP



Roadmap

- (1) Problem Formulation
- (2) **Parameter Estimation: ML**
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection



MLE (Maximum Likelihood Estimation) (1)

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

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

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

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



MLE (Maximum Likelihood Estimation) (2)

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

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

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



MLE (Maximum Likelihood Estimation) (3)

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

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

$$\iff \theta_{\text{ML}}^\top \mathbf{X}^\top \mathbf{X} = \mathbf{y}^\top \mathbf{X}$$

$$\iff \theta_{\text{ML}}^\top = \mathbf{y}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \quad (\mathbf{X}^\top \mathbf{X} \text{ is positive definite if } \text{rk}(\mathbf{X}) = D)$$

$$\iff \theta_{\text{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$



MLE with Features

- 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 $\boldsymbol{\theta} \in \mathbb{R}^K$, $\phi(\mathbf{x}) : \mathbb{R}^D \mapsto \mathbb{R}^K$:

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

- **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 $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$,

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

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

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

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

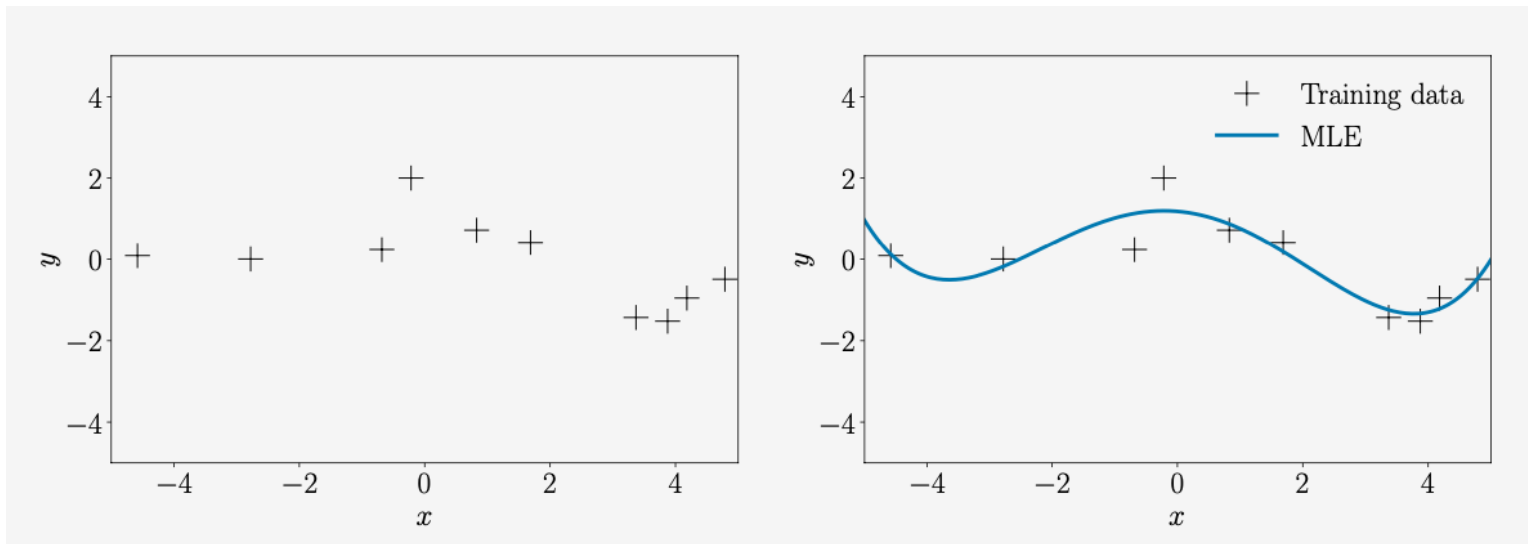
$$-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = \frac{1}{2\sigma^2} \|\mathbf{y} - \Phi\boldsymbol{\theta}\|^2 + \text{const}$$

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

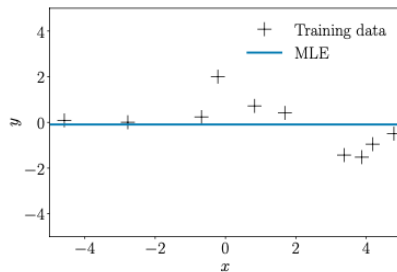


Polynomial Fit

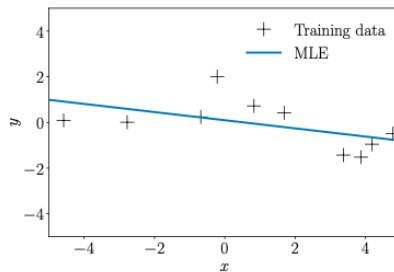
- $N = 10$ data, where $x_n \sim \mathcal{U}[-5, 5]$ and $y_n = -\sin(x_n/5) + \cos(x_n) + \epsilon$, $\epsilon \sim \mathcal{N}(0, 0.2^2)$
- Fit with polynomial with degree 4 using ML



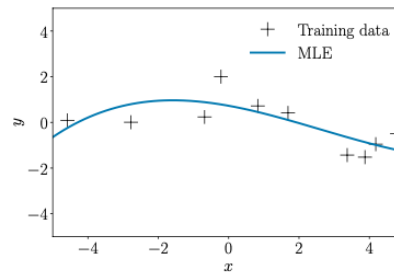
Overfitting in Linear Regression



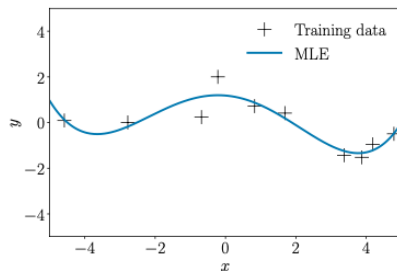
(a) $M = 0$



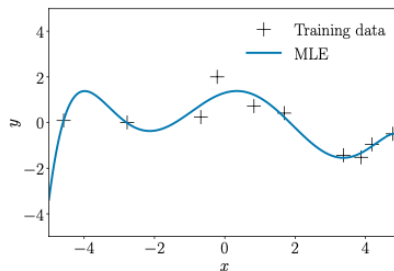
(b) $M = 1$



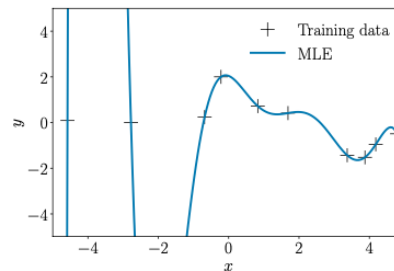
(c) $M = 3$



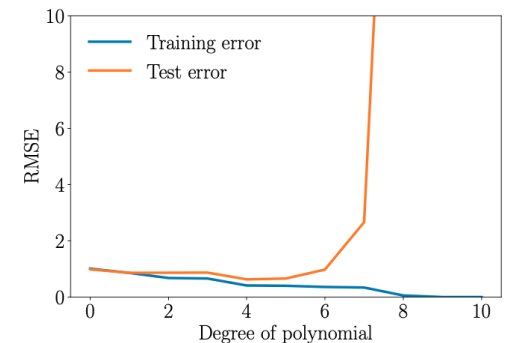
(d) $M = 4$



(e) $M = 6$



(f) $M = 9$



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



Roadmap

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) **Parameter Estimation: MAP**
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection



MAPE (Maximum A Posteriori Estimation)

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

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

- Gradient

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



MAPE for Gausssian 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 **L6(6)**
- Negative log-posterior

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

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

- Gradient

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



MAPE for Gaussssian 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 Gaussssian Prior (3)

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

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

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

- 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

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



Roadmap

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection

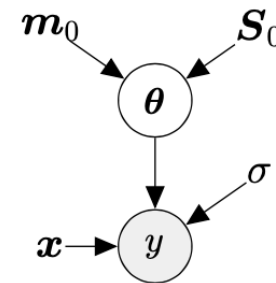


Bayesian Linear Regression

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

L8(4)

$$\begin{array}{ll} \text{prior} & p(\boldsymbol{\theta}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0) \\ \text{likelihood} & p(y|\mathbf{x}, \boldsymbol{\theta}) \sim \mathcal{N}(y | \phi^\top(\mathbf{x})\boldsymbol{\theta}, \sigma^2) \\ \text{joint} & p(y, \boldsymbol{\theta}|\mathbf{x}) = p(y | \mathbf{x}, \boldsymbol{\theta})p(\boldsymbol{\theta}) \end{array}$$



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

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

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

¹ **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

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

(Proof Sketch)

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

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



Parameter Posterior Distribution (2)

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

- 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 .

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

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



Posterior Predictions (1)

- Posterior predictive distribution

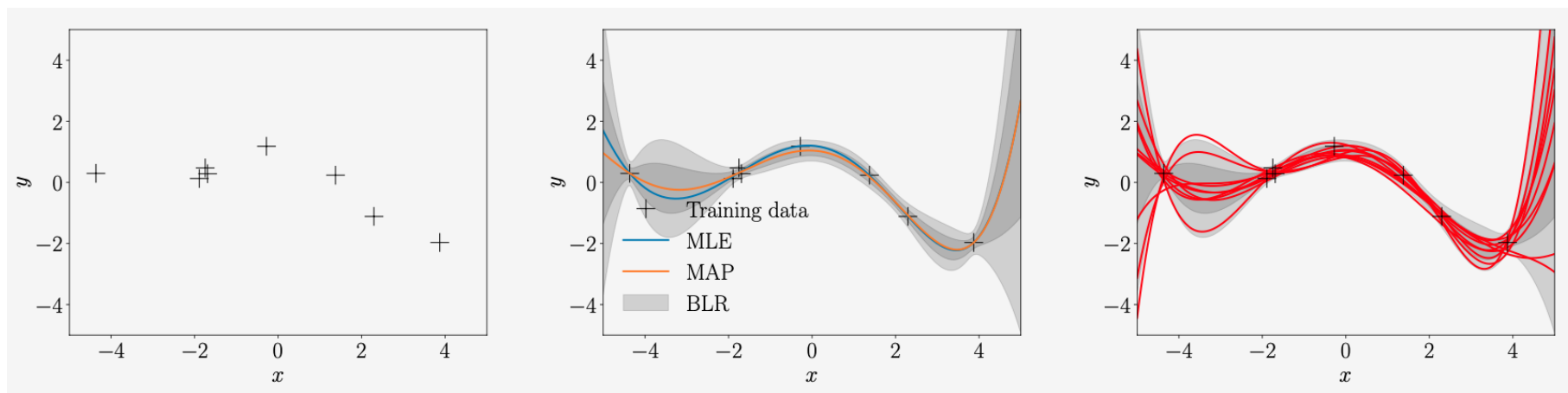
L6(5)

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

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



Posterior Predictions (2)



- BLR: Bayesian Linear Regression



Computing Marginal Likelihood

- Likelihood: $p(\mathcal{Y}|\mathcal{X}, \theta)$, Marginal likelihood:
 $p(\mathcal{Y}|\mathcal{X}) = \int p(\mathcal{Y}|\mathcal{X}, \theta)p(\theta)d\theta$
- Recall that the marginal likelihood is important for model selection via Bayes factor:

$$\text{(Posterior odds)} = \frac{\mathbb{P}(M_1 | \mathcal{D})}{\mathbb{P}(M_2 | \mathcal{D})} = \frac{\frac{\mathbb{P}(\mathcal{D}|M_1)\mathbb{P}(M_1)}{\mathbb{P}(\mathcal{D})}}{\frac{\mathbb{P}(\mathcal{D}|M_2)\mathbb{P}(M_2)}{\mathbb{P}(\mathcal{D})}} = \underbrace{\frac{\mathbb{P}(M_1)}{\mathbb{P}(M_2)}}_{\text{Prior odds}} \underbrace{\frac{\mathbb{P}(\mathcal{D} | M_1)}{\mathbb{P}(\mathcal{D} | M_2)}}_{\text{Bayes factor}}$$

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



Roadmap

- (1) Problem Formulation
- (2) Parameter Estimation: ML
- (3) Parameter Estimation: MAP
- (4) Bayesian Linear Regression
- (5) Maximum Likelihood as Orthogonal Projection



ML as Orthogonal Projection

- For $f(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\theta}_{\text{ML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \frac{\mathbf{X}^\top \mathbf{y}}{\mathbf{X}^\top \mathbf{X}} \in \mathbb{R}$

$$\mathbf{X} \boldsymbol{\theta}_{\text{ML}} = \frac{\mathbf{X} \mathbf{X}^\top}{\mathbf{X}^\top \mathbf{X}} \mathbf{y}$$

- Orthogonal projection of \mathbf{y} onto the one-dimensional subspace spanned by \mathbf{X}

- For $f(\mathbf{x}) = \boldsymbol{\phi}^\top(\mathbf{x}) \boldsymbol{\theta} + \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\theta}_{\text{ML}} = (\boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^\top \mathbf{y} = \frac{\boldsymbol{\Phi}^\top \mathbf{y}}{\boldsymbol{\Phi}^\top \boldsymbol{\Phi}} \in \mathbb{R}$

$$\boldsymbol{\Phi} \boldsymbol{\theta}_{\text{ML}} = \frac{\boldsymbol{\Phi} \boldsymbol{\Phi}^\top}{\boldsymbol{\Phi}^\top \boldsymbol{\Phi}} \mathbf{y}$$

- Orthogonal projection of \mathbf{y} onto the K -dimensional subspace spanned by columns of $\boldsymbol{\Phi}$



Summary and Other Issues (1)

- Linear regression for Gaussian likelihood and conjugate Gaussian priors. Nice analytical results and closed forms
- Other forms of likelihoods for other applications (e.g., classification)
- GLM (generalized linear model): $y = \sigma \circ f$ (σ : activation function)
 - No longer linear in θ
 - Logistic regression: $\sigma(f) = \frac{1}{1 + \exp(-f)} \in [0, 1]$ (interpreted as the probability of becoming 1)
 - Building blocks of (deep) feedforward neural nets
 - $\mathbf{y} = \sigma(\mathbf{Ax} + \mathbf{b})$. \mathbf{A} : weight matrix, \mathbf{b} : bias vector
 - K -layer deep neural nets: $\mathbf{x}_{k+1} = f_k(\mathbf{x}_k)$, $f_k(\mathbf{x}_k) = \sigma_k(\mathbf{A}_k \mathbf{x}_k + \mathbf{b}_k)$



Summary and Other Issues (2)

- Gaussian process
 - A distribution over parameters \rightarrow a distribution over functions
 - Gaussian process: distribution over functions without detouring via parameters
 - Closely related to BLR and support vector regression, also interpreted as Bayesian neural network with a single hidden layer and the infinite number of units
- Gaussian likelihood, but non-Gaussian prior
 - When $N \ll D$ (small training data)
 - Prior that enforces sparsity, e.g., Laplace prior
 - A linear regression with the Laplace prior = linear regression with LASSO (L1 regularization)



THANKS FOR YOUR ATTENTION



Discussions

