DOĞRUSAL BAĞLANIM

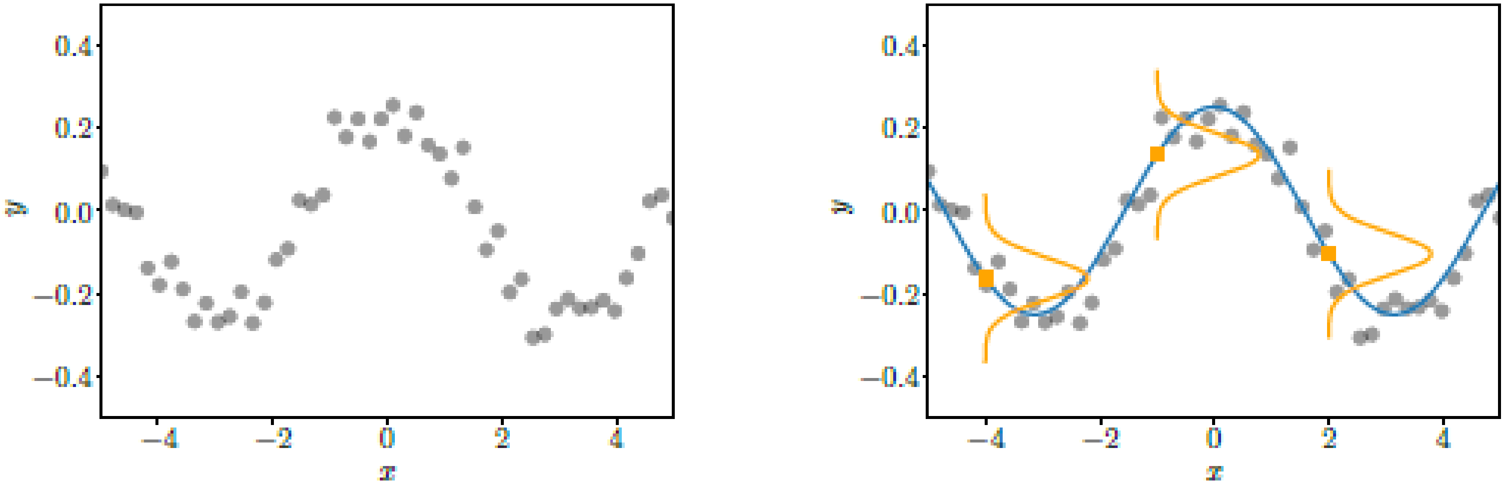
(CHAPTER 9)

Bağlanım Problemi

Gender ID	Degree	Latitude (in degrees)	Longitude (in degrees)	Age	Annual Salary (in thousands)
-1	2	51.5073	0.1290	36	89.563
-1	3	51.5074	0.1275	47	123.543
+1	1	51.5071	0.1278	26	23.989
-1	1	51.5075	0.1281	68	138.769
+1	2	51.5074	0.1278	33	113.888

$$y = f(\boldsymbol{x}) + \epsilon ,$$

$$\textit{training set } \mathcal{D} := \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$$



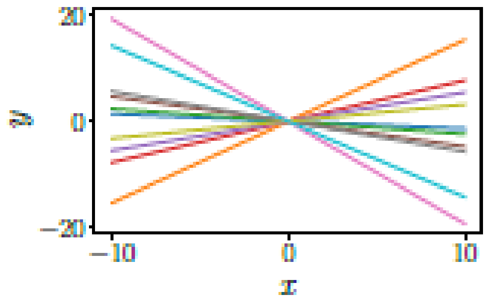
Problem Tanımı

$$y = f(\boldsymbol{x}) + \epsilon\,,$$

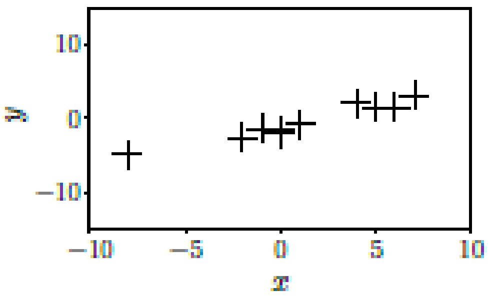
$$p(y\,|\,\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}\big(y\,|\,\boldsymbol{x}^\top\boldsymbol{\theta},\, \sigma^2\big)$$
$$\iff y = \boldsymbol{x}^\top\boldsymbol{\theta} + \epsilon\,, \quad \epsilon \sim \mathcal{N}\big(0,\, \sigma^2\big)\,,$$

- Doğrusal
- Rassal sapmalar

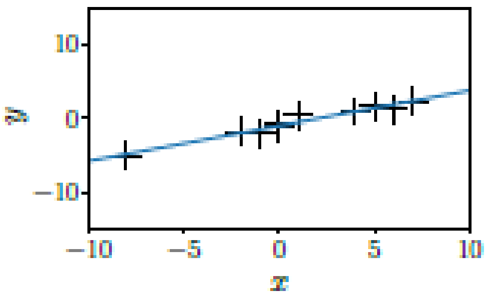
$$p(y_*\,|\,\boldsymbol{x}_*, \boldsymbol{\theta}^*) = \mathcal{N}\big(y_*\,|\,\boldsymbol{x}_*^\top\boldsymbol{\theta}^*,\, \sigma^2\big)\,.$$



(a) Example functions (straight lines) that can be described using the linear model in (9.4).



(b) Training set.



(c) Maximum likelihood estimate.

En Büyük Olabilirlik Kestirimi

$$\mathcal{X} := \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$$

$$\mathcal{Y} := \{y_1, \ldots, y_N\}$$

$$p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = p(y_1, \ldots, y_N \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{\theta})$$
$$= \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \boldsymbol{x}_n^\top \boldsymbol{\theta}, \sigma^2),$$

$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}).$$

- Negatif log-olabilirlik

$$-\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) = -\log \prod_{n=1}^{N} p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}) = -\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta}),$$

En Büyük Olabilirlik Kestirimi

$$-\sum_{n=1}^{N} \log p(y_n \mid \boldsymbol{x}_n, \boldsymbol{\theta})$$

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &:= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - \boldsymbol{x}_n^\top\boldsymbol{\theta})^2 \\ &= \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) = \frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}\|^2,\end{aligned}$$

$$\boldsymbol{X} := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^\top \in \mathbb{R}^{N \times D}$$

$$\boldsymbol{y} := [y_1, \ldots, y_N]^\top \in \mathbb{R}^N$$

- En küçük kareler

$$\boldsymbol{\theta}_{\mathrm{ML}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$$

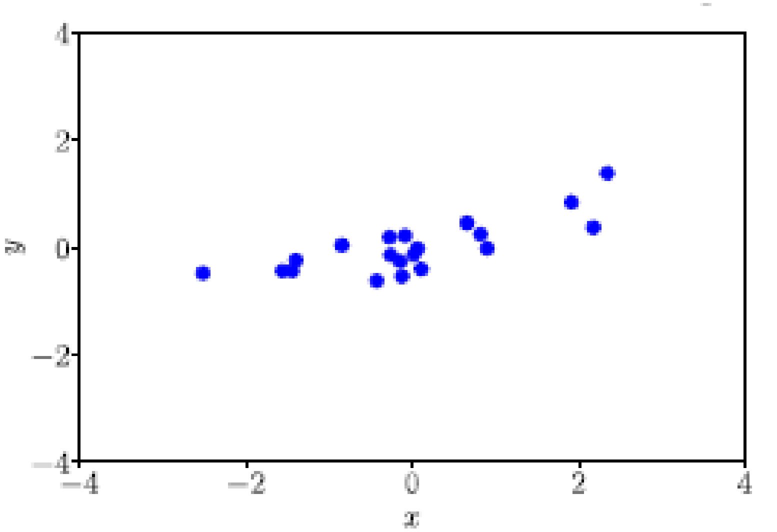
- N>= D

$$\nabla_{\boldsymbol{\theta}}^2\mathcal{L}(\boldsymbol{\theta}) = \boldsymbol{X}^\top\boldsymbol{X}$$

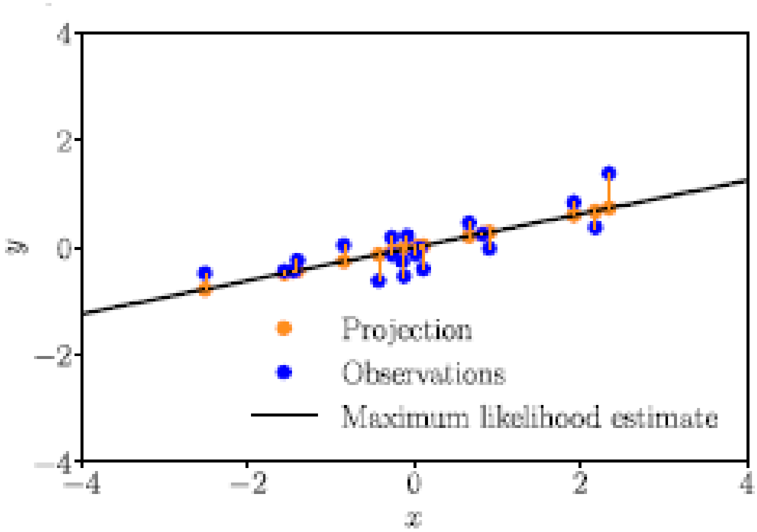
En Büyük Olabilirlik Kestirimi

- Dik izdüşüm

$$X\theta_{\mathrm{ML}} = X\frac{X^\top y}{X^\top X} = \frac{XX^\top}{X^\top X}y\,,$$



(a) Regression dataset consisting of noisy observations y_n (blue) of function values $f(x_n)$ at input locations x_n.



(b) The orange dots are the projections of the noisy observations (blue dots) onto the line $\theta_{\mathrm{ML}}x$. The maximum likelihood solution to a linear regression problem finds a subspace (line) onto which the overall projection error (orange lines) of the observations is minimized.

En Büyük Olabilirlik Kestirimi

$$y = \boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta} + \epsilon = \sum_{k=0}^{K-1} \theta_k \phi_k(\boldsymbol{x}) + \epsilon\,,$$

$$\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^K$$

$$\phi_k : \mathbb{R}^D \to \mathbb{R}$$

- Örnek: Polinomsal bağlanım

$$\boldsymbol{\phi}(x) = \begin{bmatrix} \phi_0(x) \\ \phi_1(x) \\ \vdots \\ \phi_{K-1}(x) \end{bmatrix} = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ \vdots \\ x^{K-1} \end{bmatrix} \in \mathbb{R}^K$$

- Özellikler matrisi

$$\boldsymbol{\Phi} := \begin{bmatrix} \boldsymbol{\phi}^\top(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}^\top(\boldsymbol{x}_N) \end{bmatrix} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \cdots & \phi_{K-1}(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \cdots & \phi_{K-1}(\boldsymbol{x}_2) \\ \vdots & & \vdots \\ \phi_0(\boldsymbol{x}_N) & \cdots & \phi_{K-1}(\boldsymbol{x}_N) \end{bmatrix} \in \mathbb{R}^{N\times K}\,,$$

where $\Phi_{ij} = \phi_j(\boldsymbol{x}_i)$ and $\phi_j : \mathbb{R}^D \to \mathbb{R}$.

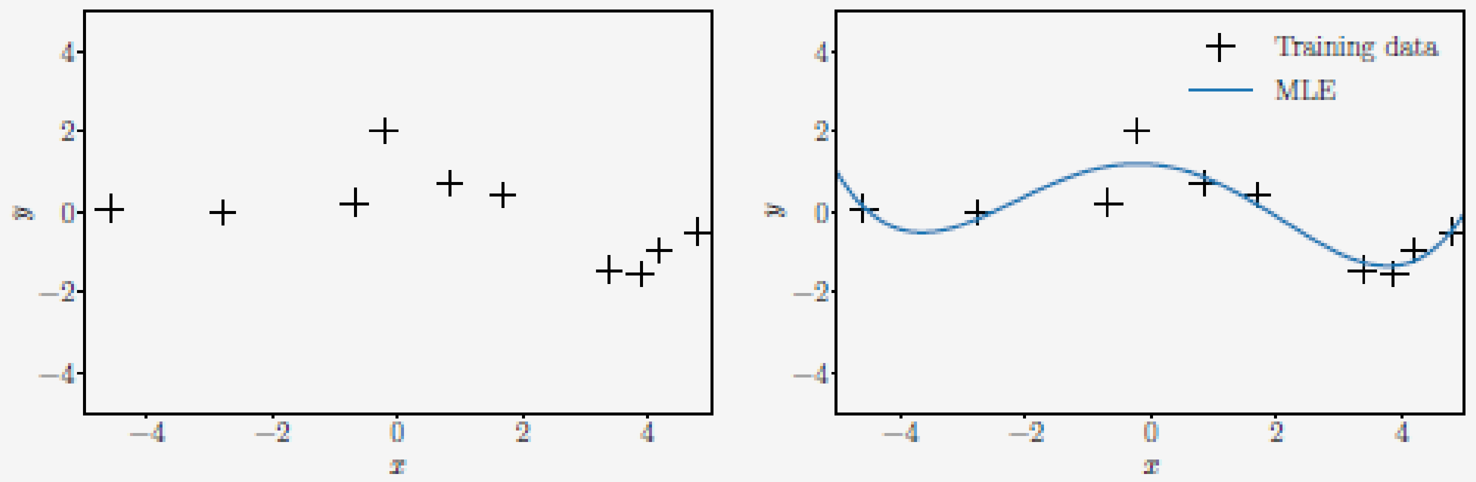
$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_N & x_N^2 \end{bmatrix}.$$

En Büyük Olabilirlik Kestirimi

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

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

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



(a) Regression dataset.

(b) Polynomial of degree 4 determined by maximum likelihood estimation.

En Büyük Olabilirlik Kestirimi

$$\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}, \sigma^2) = \sum_{n=1}^{N} \log \mathcal{N}\left(y_n \mid \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta},\, \sigma^2\right)$$

$$= \sum_{n=1}^{N} \left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y_n - \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta})^2\right)$$

$$= -\frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\underbrace{\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta})^2}_{=:s} + \text{const}.$$

$$\frac{\partial \log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}, \sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}s = 0$$

$$\iff \frac{N}{2\sigma^2} = \frac{s}{2\sigma^4}$$

$$\sigma^2_{\mathrm{ML}} = \frac{s}{N} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\phi}^\top(\boldsymbol{x}_n)\boldsymbol{\theta})^2 .$$

Model Başarısı

- RMSE

$$\sqrt{\frac{1}{N}\left\|y-\Phi\theta\right\|^2}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n-\phi^\top(x_n)\theta)^2}\,,$$

- Aşırı uyumlanma

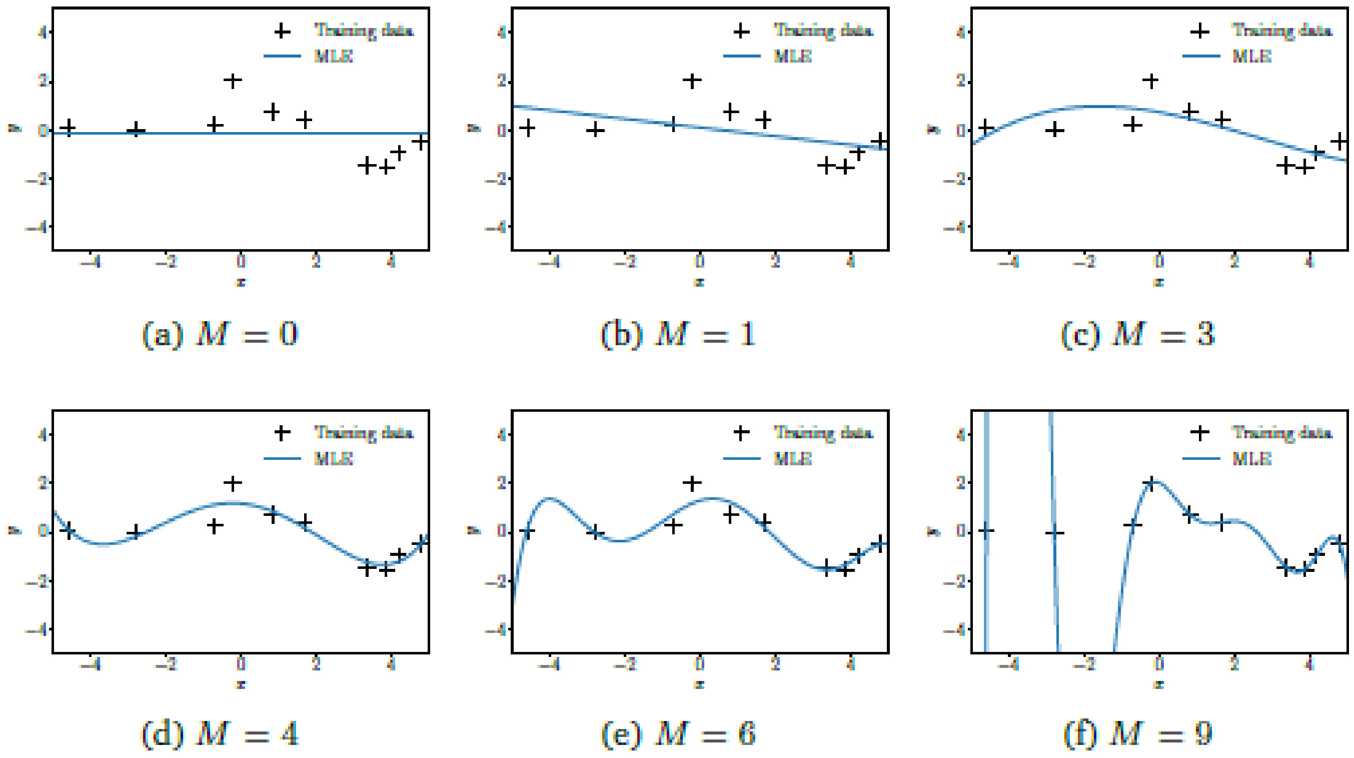
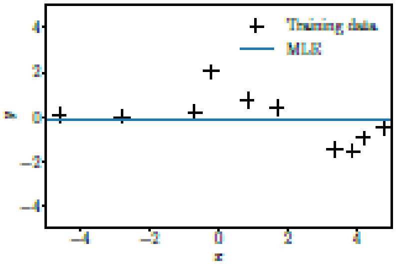


Figure 9.5 Maximum likelihood fits for different polynomial degrees M.

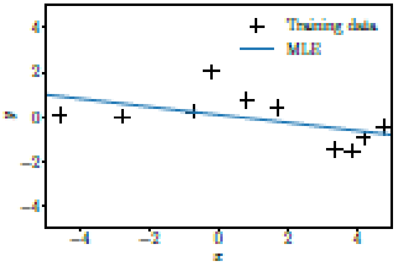
Model Başarısı

- RMSE

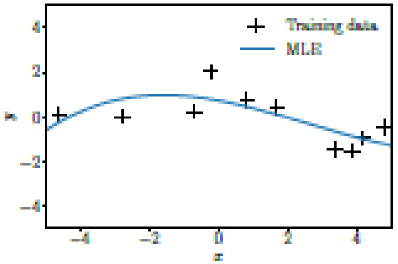
$$\sqrt{\frac{1}{N}\left\|\boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\theta}\right\|^{2}}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(y_n-\boldsymbol{\phi}^{\top}(\boldsymbol{x}_n)\boldsymbol{\theta})^2}\,,$$



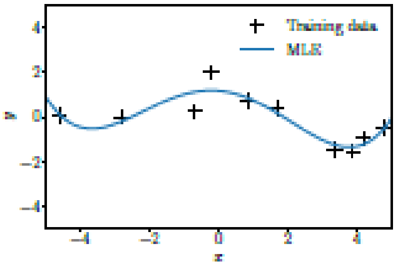
(a) $M = 0$



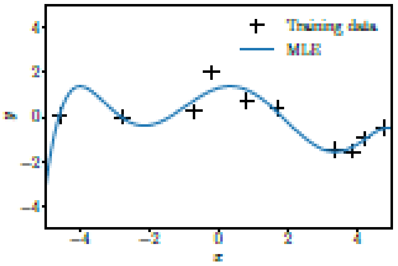
(b) $M = 1$



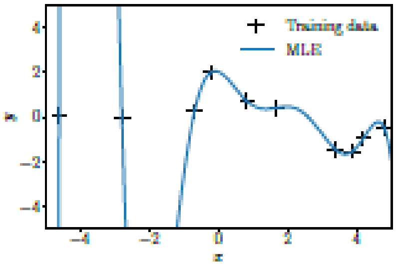
(c) $M = 3$



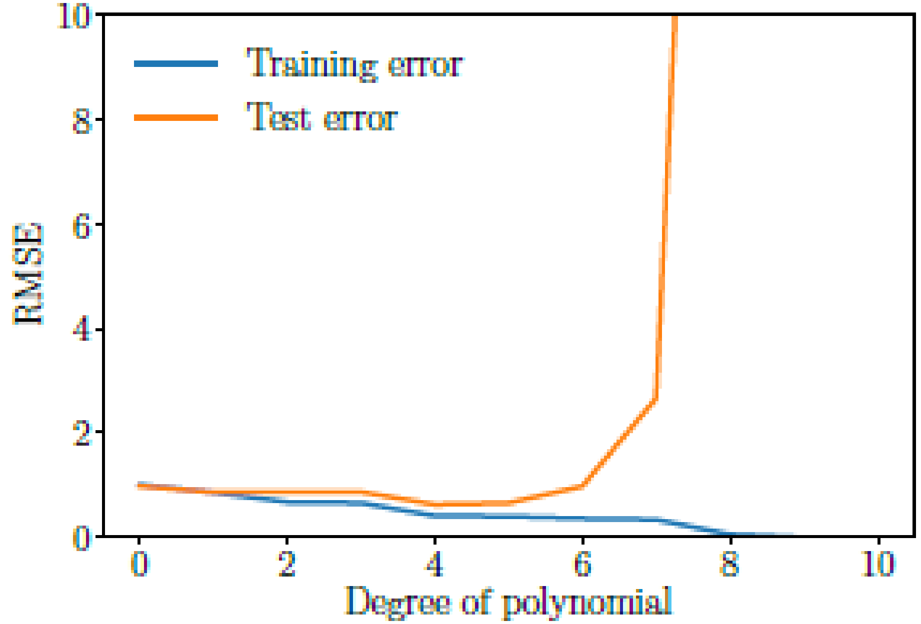
(d) $M = 4$



(e) $M = 6$



(f) $M = 9$



En Büyük Sonsal Olasılık Kestirimi

- Öncül dağılım $p(\boldsymbol{\theta})$

$$p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathcal{Y} \mid \mathcal{X})} .$$

$$\log p(\boldsymbol{\theta} \mid \mathcal{X}, \mathcal{Y}) = \log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) + \text{const} ,$$

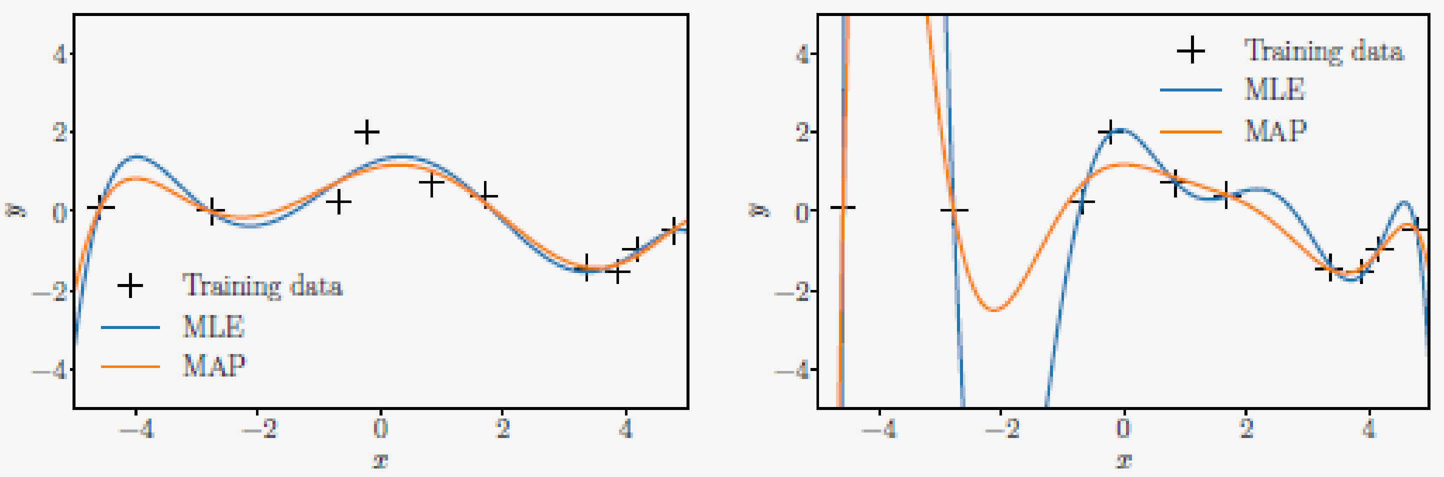
$$\boldsymbol{\theta}_{\text{MAP}} \in \arg\min_{\boldsymbol{\theta}} \{ -\log p(\mathcal{Y} \mid \mathcal{X}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \} .$$

En Büyük Sonsal Olasılık Kestirimi

Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2\boldsymbol{I})$

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

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



(a) Polynomials of degree 6. (b) Polynomials of degree 8.

En Büyük Sonsal Olasılık Kestirimi

Gaussian prior $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0}, b^2\boldsymbol{I})$

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

$$\|\boldsymbol{y} - \boldsymbol{\Phi\theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_2^2$$

- Düzenlileştirilmiş en küçük kareler
- Norm seçimi

Bayesçi Doğrusal Regresyon

$$\begin{aligned} &\text{prior} && p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{m}_0,\, \boldsymbol{S}_0)\,, \\ &\text{likelihood} && p(y\,|\,\boldsymbol{x}, \boldsymbol{\theta}) = \mathcal{N}(y\,|\,\boldsymbol{\phi}^\top(\boldsymbol{x})\boldsymbol{\theta},\, \sigma^2)\,, \end{aligned} \tag{9.35}$$

Theorem 9.1 (Parameter Posterior). *In our model (9.35), the parameter posterior (9.41) can be computed in closed form as*

$$p(\boldsymbol{\theta}\,|\,\mathcal{X}, \mathcal{Y}) = \mathcal{N}(\boldsymbol{\theta}\,|\,\boldsymbol{m}_N,\, \boldsymbol{S}_N)\,, \tag{9.43a}$$

$$\boldsymbol{S}_N = (\boldsymbol{S}_0^{-1} + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\,, \tag{9.43b}$$

$$\boldsymbol{m}_N = \boldsymbol{S}_N(\boldsymbol{S}_0^{-1}\boldsymbol{m}_0 + \sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{y})\,, \tag{9.43c}$$

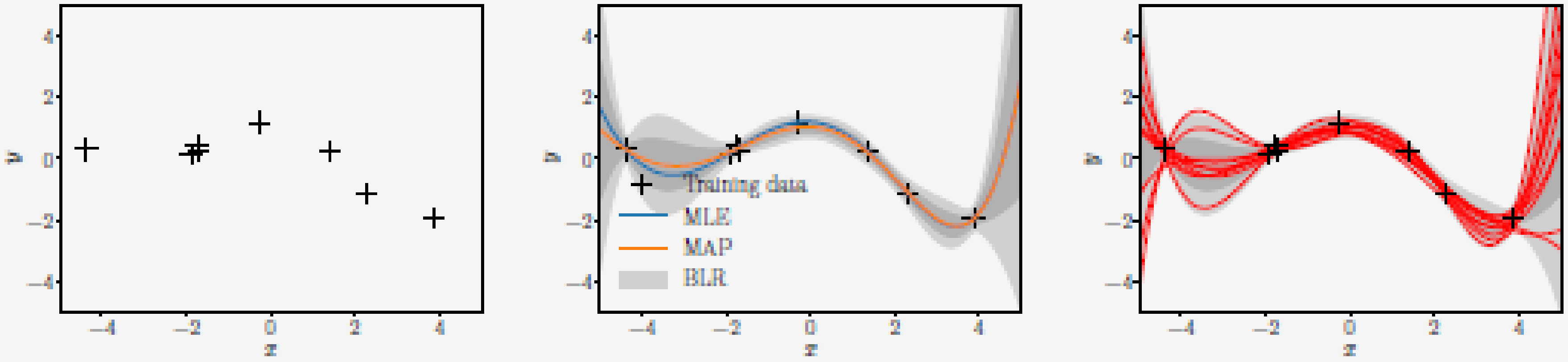
where the subscript N indicates the size of the training set.

Sonsal Tahmin Dağılımı

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

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{x}_*) \,|\, \mathcal{X}, \mathcal{Y}] &= \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta} \,|\, \mathcal{X}, \mathcal{Y}] = \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \,|\, \mathcal{X}, \mathcal{Y}] \\
&= \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{m}_N = \boldsymbol{m}_N^\top\boldsymbol{\phi}(\boldsymbol{x}_*) , \\
\mathbb{V}_{\boldsymbol{\theta}}[f(\boldsymbol{x}_*) \,|\, \mathcal{X}, \mathcal{Y}] &= \mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{\theta} \,|\, \mathcal{X}, \mathcal{Y}] \\
&= \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\mathbb{V}_{\boldsymbol{\theta}}[\boldsymbol{\theta} \,|\, \mathcal{X}, \mathcal{Y}]\boldsymbol{\phi}(\boldsymbol{x}_*) \\
&= \boldsymbol{\phi}^\top(\boldsymbol{x}_*)\boldsymbol{S}_N\boldsymbol{\phi}(\boldsymbol{x}_*) .
\end{aligned}$$

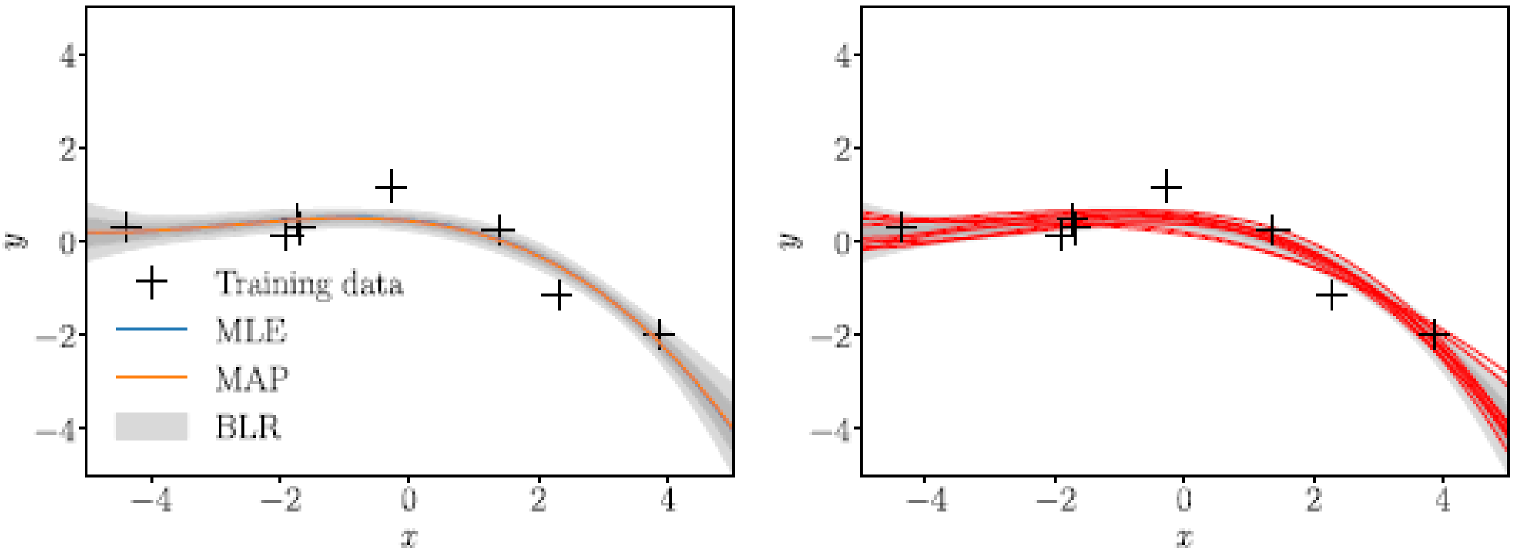
Fonksiyonların Sonsal Dağılımı



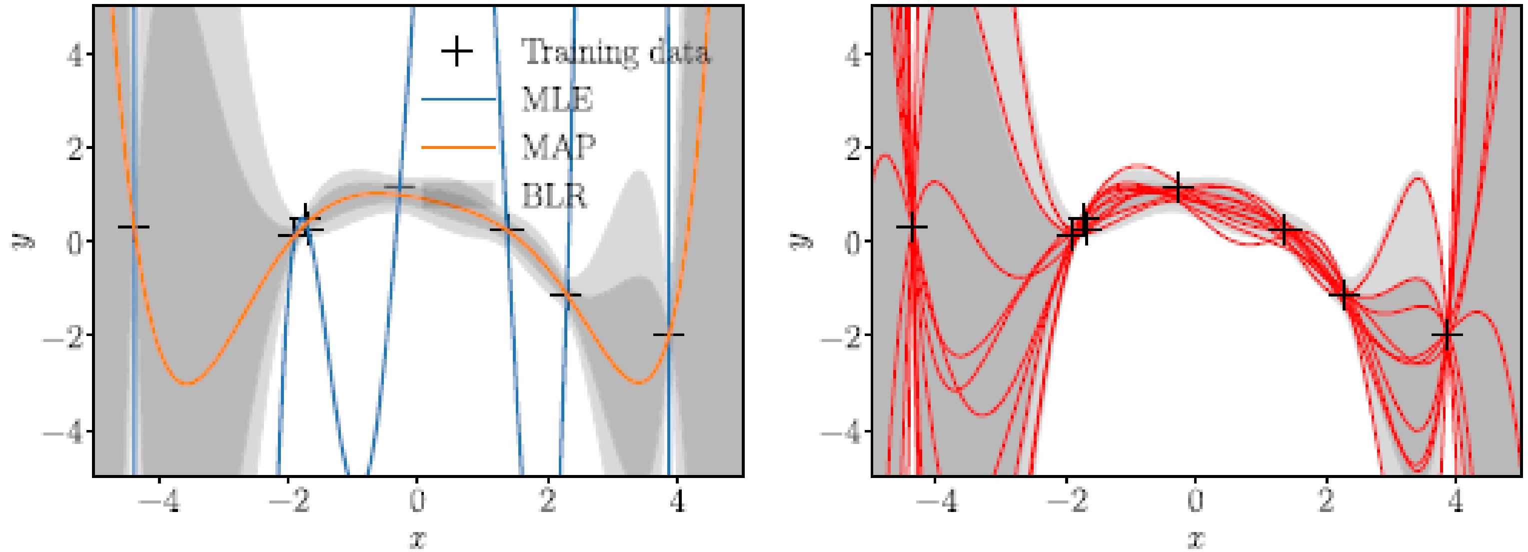
(a) Training data.

(b) Posterior over functions represented by the marginal uncertainties (shaded) showing the 67% and 95% predictive confidence bounds, the maximum likelihood estimate (MLE) and the MAP estimate (MAP), the latter of which is identical to the posterior mean function.

(c) Samples from the posterior over functions, which are induced by the samples from the parameter posterior.



(a) Posterior distribution for polynomials of degree $M = 3$ (left) and samples from the posterior over functions (right).



(c) Posterior distribution for polynomials of degree $M = 7$ (left) and samples from the posterior over functions (right).