Regression Analysis I

Jin Young Choi
Seoul National University

Outline

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression

LINEAR REGRESSION

JIN YOUNG CHOI

ECE, SEOUL NATIONAL UNIVERSITY

http://3.droppdf.com/files/pjxkI/regression-analysis-by-example-5th-edition.pdf
https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial

Regression Analysis

- For independent random variable X, and dependent random variable Y, assume they have a functional correlation between them, i.e.

$$Y = f(X)$$

- Regression: a process to find a parametric model $\hat{f}$ that gives the best fit of f for the observed samples

$$Y = \hat{f}(X) + \epsilon, \quad X\text{: predictor r.v., } Y\text{: response r.v.}$$

- Assume $E(\epsilon) = 0$, var$(\epsilon) = \sigma^2$, then $E(Y|x) = \hat{f}(x)$ for an observed non-random value x
- $\hat{f}$ can be estimated from the sample pairs $\{(y_i, x_i)|\ i = 1, 2, \cdots, n\}$

$$y_i = \hat{f}(x_i) + \epsilon_i, \ \ i = 1, \ \cdots, \ n,$$

where ϵ_i are i.i.d. zero mean and variance σ^2

Simple Linear Regression

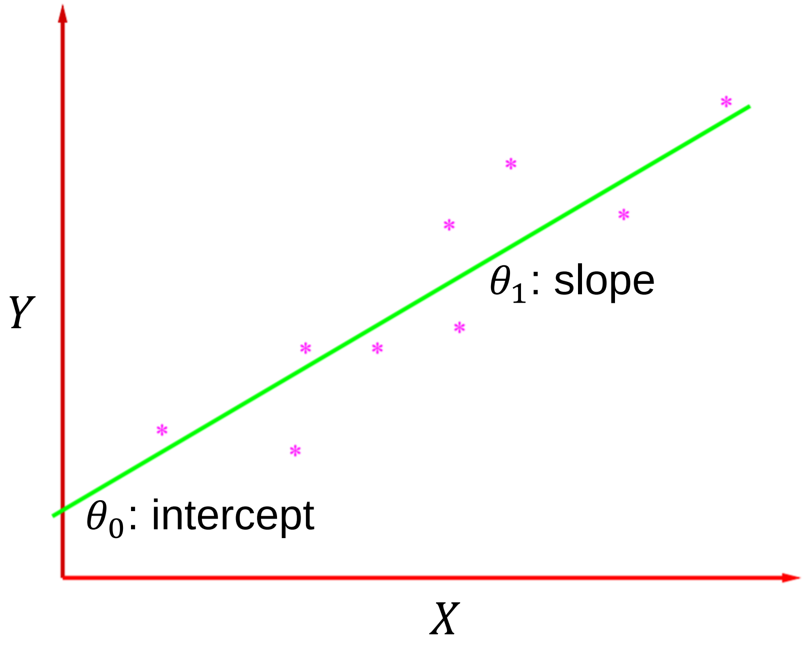
- Simple linear regression model

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_i + \epsilon_i, \; i = 1, \cdots, n,$$

where θ_0: intercept, θ_1: slope

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
⋮	⋮	⋮
n	y_n	x_n



Simple Linear Regression

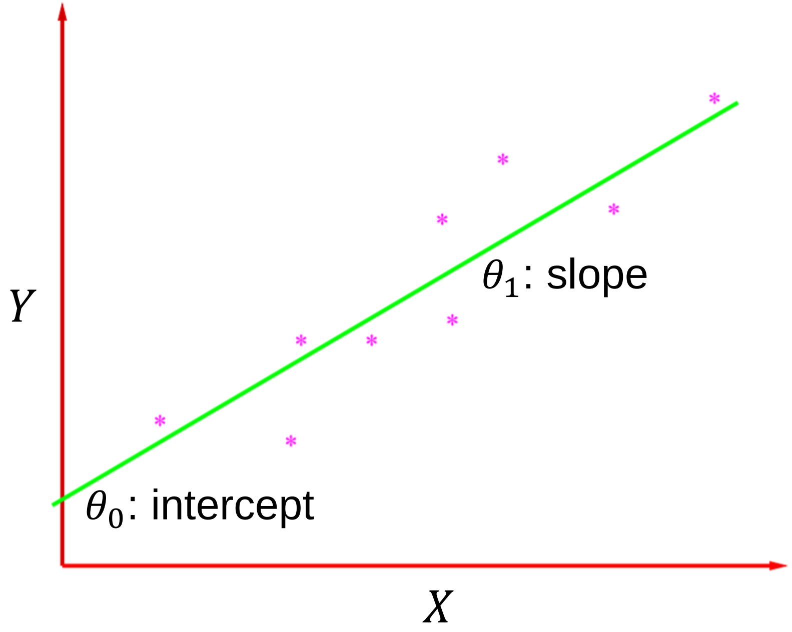
- Correlation of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\mathrm{Cov}(Y, X) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})\,(x_i - \bar{x})$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Observation Number	Response Y	Predictor X
1	y_1	x_1
2	y_2	x_2
3	y_3	x_3
⋮	⋮	⋮
n	y_n	x_n



Simple Linear Regression

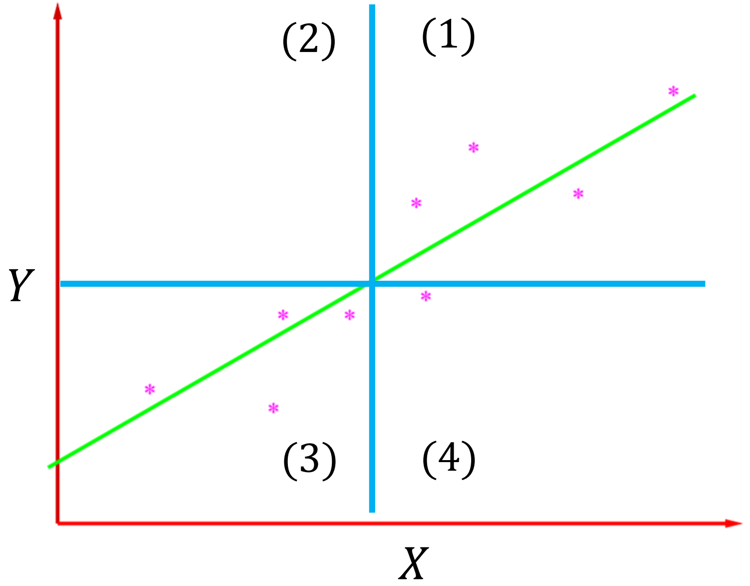
- Correlation of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\mathrm{Cov}(Y,X) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})\,(x_i - \bar{x})$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Q	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
(1)	$+$	$+$	$+$
(2)	$+$	$-$	$-$
(3)	$-$	$-$	$+$
(4)	$-$	$+$	$-$



$\theta_1 \geq 0 \longrightarrow \mathrm{Cov}(Y,X) \geq 0$

$\theta_1 < 0 \longrightarrow \mathrm{Cov}(Y,X) < 0$

Simple Linear Regression

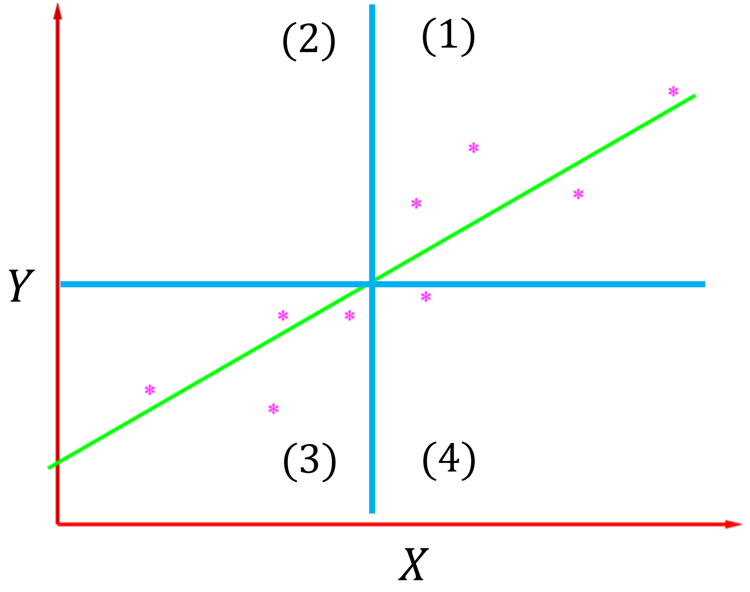
- Correlation Coefficient of Y & X

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$\rho(Y,X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\bar{y}}{\sigma_y}\right)\left(\frac{x_i-\bar{x}}{\sigma_x}\right)$$

where $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2$, $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Q	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})(x_i - \bar{x})$
(1)	$+$	$+$	$+$
(2)	$+$	$-$	$-$
(3)	$-$	$-$	$+$
(4)	$-$	$+$	$-$



$\theta_1 \geq 0 \longrightarrow 1 \geq \rho\,(Y,X) \geq 0$

$\theta_1 < 0 \longrightarrow -1 \leq \rho\,(Y,X) < 0$

Parameter Estimation

- ## Least Squares Estimation

Parameters are estimated by maximum likelihood estimation (MLE)

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i,\ i = 1,\ \cdots,\ n,\quad \epsilon_i \sim N(0, \sigma^2)$$

MLE:

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

LSE:

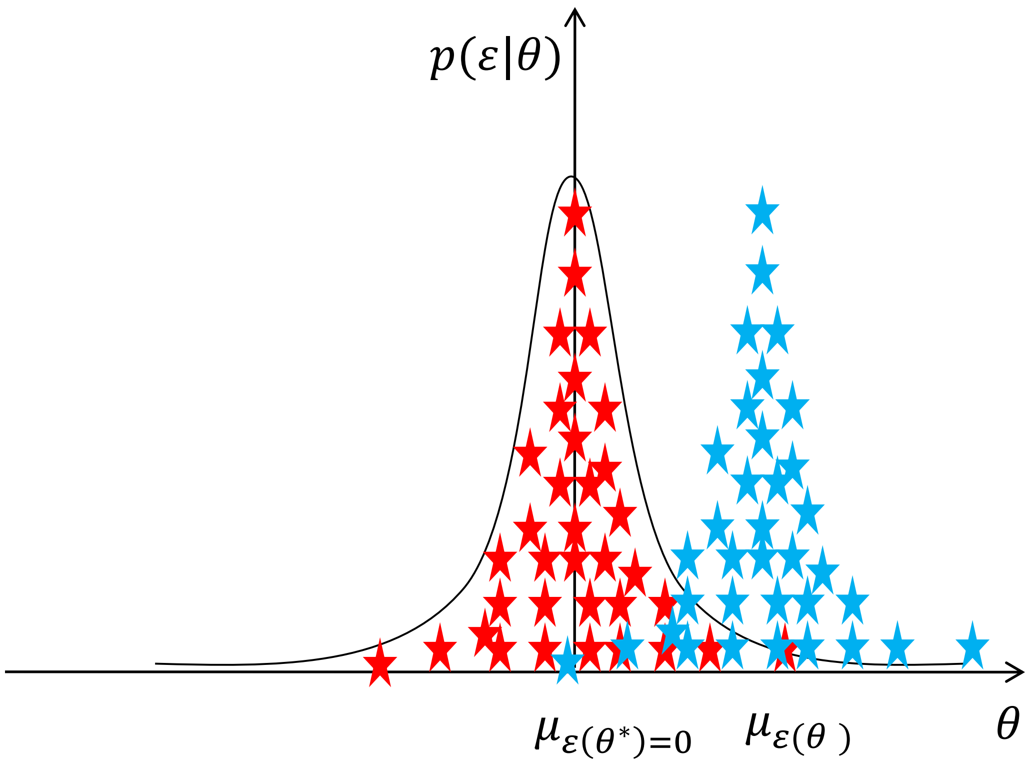
minimizing $S(\theta_0,\ \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$.

Solution:

by $\partial S/\partial\theta_0 = 0, \partial S/\partial\theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

Maximum Likelihood Estimation

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, p(\{\varepsilon_i\}|\theta) \qquad \varepsilon(\theta) = Y - \theta X$$



$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_n^T \end{bmatrix} \theta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}, \ \Phi_k = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{np} \end{bmatrix}$$

$$y_i = \phi_i^T \theta + \epsilon_i$$

$$y_i = \theta_0 + \theta_1 \phi_{i1} + \theta_2 \phi_{i2} + \cdots + \theta_p \phi_{i(p-1)} + \epsilon_i,$$

$$i = 1, \cdots, n$$

Parameter Estimation

- Least Squares Estimation

Parameters are estimated by maximum likelihood estimation (MLE)

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i,\ i = 1,\ \cdots,\ n,\quad \epsilon_i \sim N(0, \sigma^2)$$

MLE:

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmax}} \ln \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{{\epsilon_i}^2}{2\sigma^2})$$

$$(\hat{\theta}_0,\ \hat{\theta}_1) = \underset{(\theta_0,\theta_1)}{\operatorname{argmin}} \sum_{i=1}^{n} {\epsilon_i}^2 = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

LSE:

minimizing $\quad S(\theta_0,\ \theta_1) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$.

Solution:

by $\partial S / \partial \theta_0 = 0, \partial S / \partial \theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

Parameter Estimation

- Least Squares Estimation

$$\epsilon_i = y_i - \theta_0 + \theta_1 x_i, \qquad i = 1, \quad \cdots, \quad n.$$

LSE:

$$\left(\hat{\theta}_0,\ \hat{\theta}_1\right) = \underset{(\theta_0, \theta_1)}{\operatorname{argmin}}\ S(\theta_0,\ \theta_1) = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2 .$$

Solution:

by $\partial S/\partial\theta_0 = 0, \partial S/\partial\theta_1 = 0$ at $\hat{\theta}_0$ & $\hat{\theta}_1$,

$$\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right) = 0, \quad \rightarrow \boxed{\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}}$$

$$\sum_{i=1}^{n}\left(y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i\right)x_i = 0, \rightarrow \sum_{i=1}^{n}\left(y_i - \bar{y} - \hat{\theta}_1(x_i - \bar{x})\right)(x_i - \bar{x} + \bar{x}) = 0,$$

$$\rightarrow \sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\theta}_1 \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \rightarrow \boxed{\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

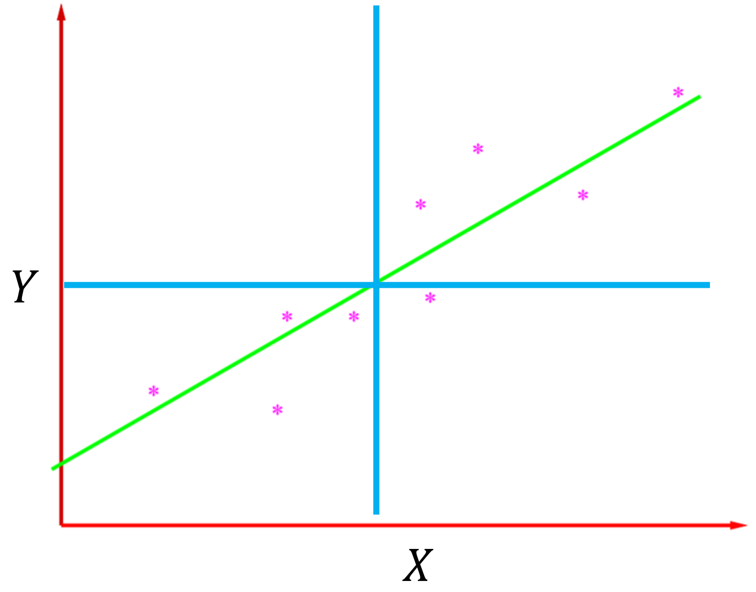
Parameter Estimation

- Least squares regression line

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X.$$

Fitted values:

$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i, \qquad i = 1, \cdots, n.$$



Error to the i-th observation:

$$e_i = y_i - \hat{y}_i, \;\; i = 1, \cdots, n.$$

Alternative formula for $\hat{\theta}_1$:

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{Cov(Y,X)}{Var(X)} = \frac{\rho(Y,X)\sigma_x\sigma_y}{\sigma_x^2} = \rho(Y,X)\frac{\sigma_y}{\sigma_x}$$

$\rightarrow$ slope has the same sign with the correlation ($\rho(Y,X)$; covariance)

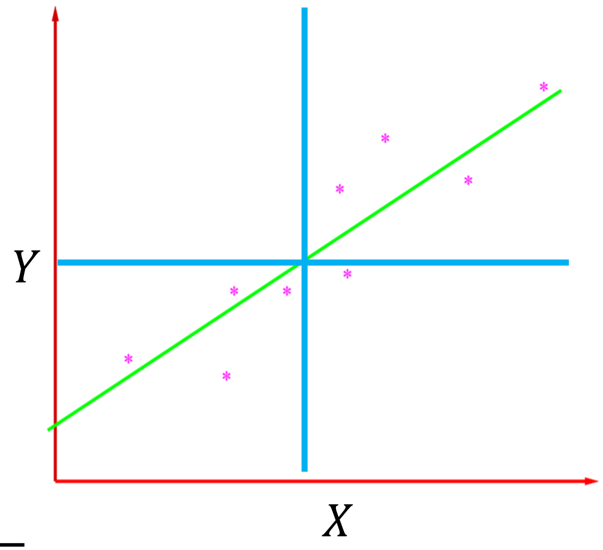
Measuring the Quality of Fit

- Original Model:

$$Y = \theta_0 + \theta_1 X + \epsilon.$$

Least squares regression line:

$$\hat{Y} = \hat{\theta}_0 + \hat{\theta}_1 X.$$



- Correlation between Y & $\hat{Y}$:

$$\rho(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)}}$$

Note that $\rho(Y, \hat{Y})$ can not be negative. Why?

Note that $\rho(Y, \hat{Y}) = 1$ implies the perfect fit.

Measuring the Quality of Fit

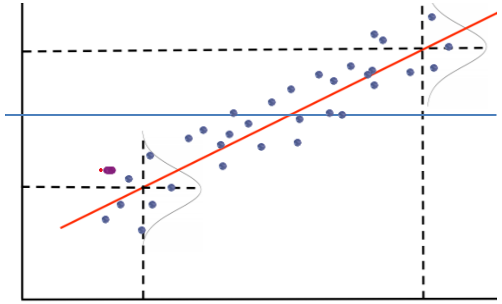
- Goodness-of-fit index:

  SST: $\sum_{i=1}^{n}(y_i - \bar{y})^2$, SST: Total sum of squares

  SSR: $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, SSR: Regression (explained) sum of squares

  SSE: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, SSE: Residual (error) sum of squares

- Interpretation:



y_i	$=$	$\hat{y}_i$	$+$	$y_i - \hat{y}_i$
Observed	=	Fit	+	Error
$y_i - \bar{y}$	$=$	$\hat{y}_i - \bar{y}$	$+$	$y_i - \hat{y}_i$
Deviation		Deviation to Fit		Residual

$$SST = SSR + SSE \quad \because \sum_{i=1}^{n}(\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0 \text{ [1]}$$

- R^2: Coefficient of determination

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$ ($R = 1$ implies the perfect fit)

[1] https://math.stackexchange.com/questions/709419/prove-sst-ssessr

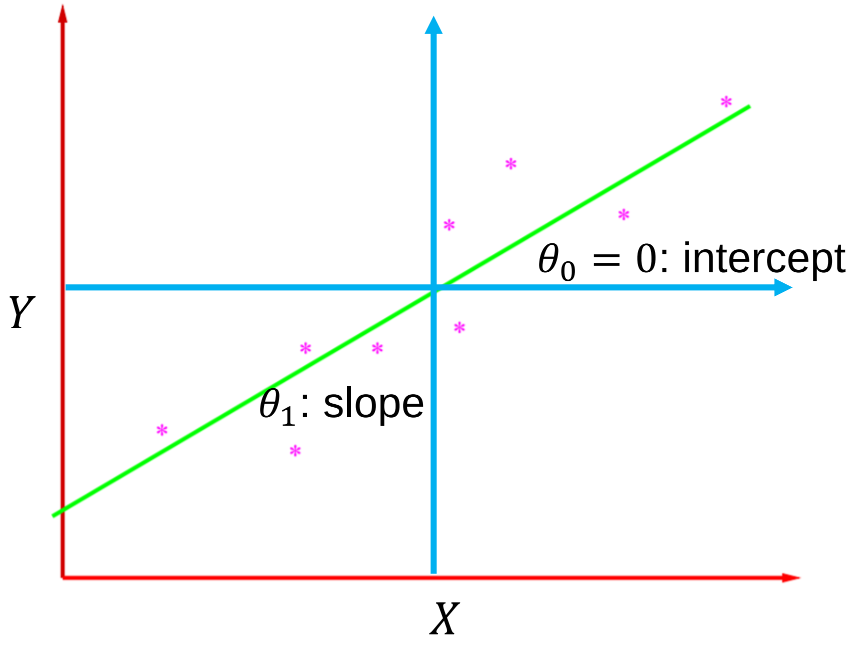
Regression Line through Origin

- Simple linear regression model

$$Y = \theta_0 + \theta_1 X + \epsilon$$

$$Y = \theta_1 X + \epsilon,$$ *no-intercept* model, $\bar{y} = \bar{x} = 0$

Observation Number	Response Y	Predictor X
1	$y_1 - \bar{y}$	$x_1 - \bar{x}$
2	$y_2 - \bar{y}$	$x_2 - \bar{x}$
3	$y_3 - \bar{y}$	$x_3 - \bar{x}$
⋮	⋮	⋮
n	$y_n - \bar{y}$	$x_n - \bar{x}$



Regression Line through Origin

- *no-intercept* model

$$y_i = \theta_1 x_i + \epsilon,$$
$$\hat{y}_i = \hat{\theta}_1 x_i,\ \ i = 1,\cdots,n$$
$$e_i = y_i - \hat{y}_i.$$

$$Cov(Y,X) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})\ (x_i - \bar{x}) \rightarrow Cov(Y,X) = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$$

$$\rho(Y,X) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i x_i}{\sigma_y \sigma_x}\ ,\quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} {y_i}^2\ , \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2$$

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \rightarrow \hat{\theta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} {x_i}^2} = \frac{Cov(Y,X)}{\sigma_x^2} = \rho(Y,X)\frac{\sigma_y}{\sigma_x}$$

$$R^2 = \frac{\sum_{i=1}^{n} {\hat{y}_i}^2}{\sum_{i=1}^{n} {y_i}^2} = 1 - \frac{\sum_{i=1}^{n} {e_i}^2}{\sum_{i=1}^{n} {y_i}^2}$$

Multivariate Linear Regression

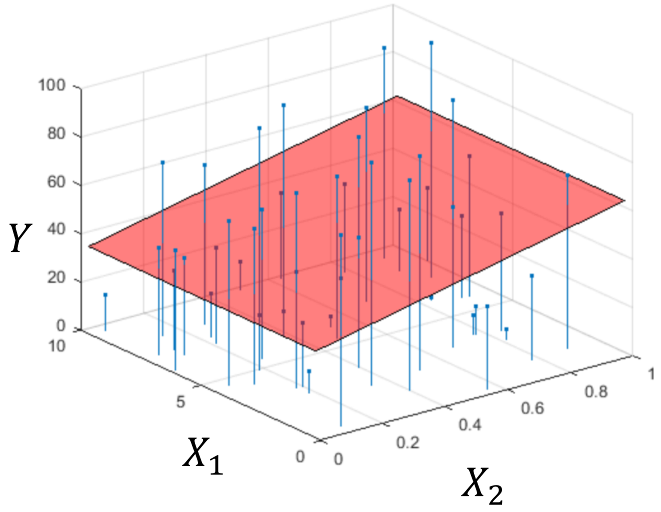
- Multivariate linear regression model: p predictor (explanatory) variables

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i,\ \ i = 1,\ \cdots,\ n,$$

where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector $(ex.;\ y = w^T x + b)$

i	Y	Predictor			
		X_1	X_2	$\cdots$	X_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
2	y_2	x_{21}	x_{22}	$\cdots$	x_{2p}
3	y_3	x_{31}	x_{32}	$\cdots$	x_{3p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}



Multivariate Linear Regression

- Multivariate linear regression model: p predictor (explanatory) variables

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i, \; i = 1, \; \cdots, \; n,$$

where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector

- Fitted model by LSE: $n - p - 1$; degree of freedom (df); $p + 1$; $\#\ of\ estimated\ parameters$

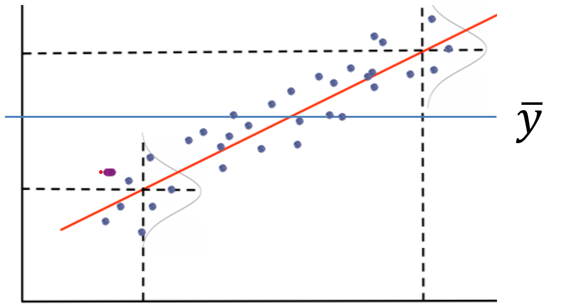
$$\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_{i1} + \hat{\theta}_2 x_{i2} + \cdots + \hat{\theta}_p x_{ip}, i = 1, \cdots, n,$$

$$e_i = y_i - \hat{y}_i.$$

- Measuring Quality of Fit:

$$\rho(Y, \hat{Y}) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)}}$$

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$



- Adjusted R^2 : $R_a^2 = 1 - \frac{1/(n-p-1) \sum_{i=1}^{n} e_i^2}{1/(n-1) \sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{n-1}{n-p-1}(1 - R^2)$

Multivariate Linear Regression

- Tests of Hypotheses for Multivariate linear model

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_p X_p + \epsilon$$

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \cdots + \theta_p x_{ip} + \epsilon_i, \;\; i = 1, \;\cdots, \; n,$$

where θ_0: intercept, $(\theta_1, \theta_2, \cdots, \theta_p)$: normal vector

- Hypotheses: H_0: Reduced model (RM), H_1: Full model (FM)
  1. All the regression coefficients associated with the predictor variables are zero.
  2. Some of the regression coefficients are zero.
  3. Some of the regression coefficients are equal to each other.
  4. The regression parameters satisfy certain specified constraints (ex. $|\theta_i| \leq \alpha$).
- Sum of Squares: $SSE(RM) \geq SSE(FM)$

$$SSE(FM) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$SSE(RM) = \sum_{i=1}^{n} (y_i - \hat{y}_i^*)^2$$

- F-test: $F = \dfrac{[SSE(RM) - SSE(FM)]/(p+1-k)}{SSE(FM)/(n-p-1)}$ (F is large $\rightarrow$ RM is inadequate[†])

[†] The critical values are given in Table A.4 and A.5 in "Regression Analysis by Example", S. Chatterjee et.al., Wiley.

NONLINEAR REGRESSION

JIN YOUNG CHOI

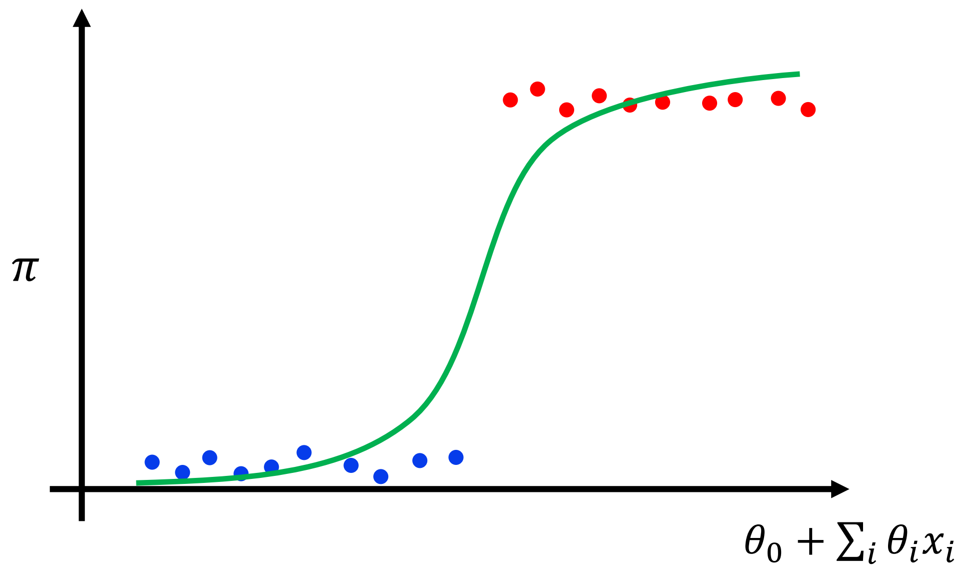
ECE, SEOUL NATIONAL UNIVERSITY

https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial

Logistic Regression

- Logistic response function representing the relation between the probability π and $X_1, X_2, \cdots, X_p$

$$\pi = p\left(Y = 1 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{\exp(\theta_0+\theta_1 x_1+\cdots+\theta_p x_p)}{1+\exp(\theta_0+\theta_1 x_1+\cdots+\theta_p x_p)}$$



Logistic Regression

- Logistic response function

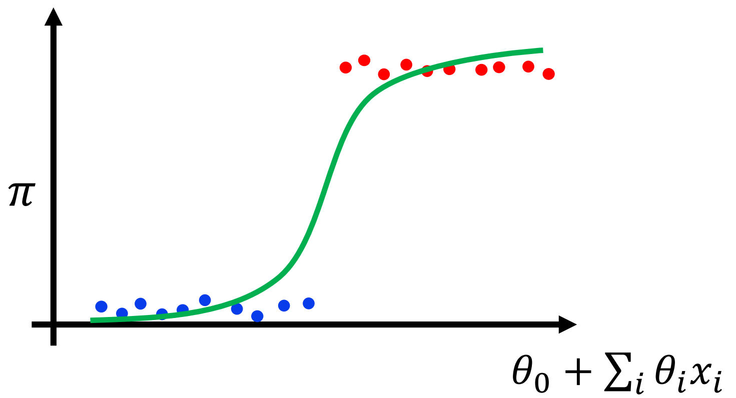
$$\pi(X_1 = x_1, \cdots, X_p = x_p) = p\left(Y = 1 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{\exp(\theta_0+\theta_1 x_1+\cdots+\theta_p x_p)}{1+\exp(\theta_0+\theta_1 x_1+\cdots+\theta_p x_p)}$$

$$1 - \pi(X_1 = x_1, \cdots, X_p = x_p) = p\left(Y = 0 \middle| X_1 = x_1, \cdots, X_p = x_p\right) = \frac{1}{1+\exp(\theta_0+\theta_1 x_1+\cdots+\theta_p x_p)}$$

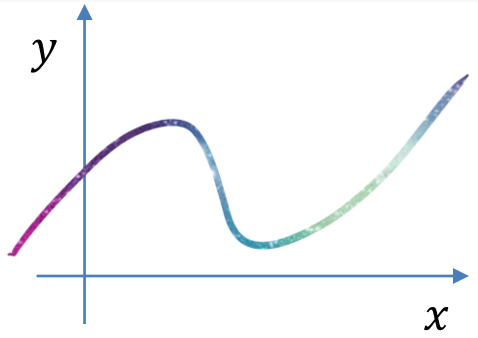
$$\frac{\pi}{1-\pi} = \exp(\theta_0 + \theta_1 x_1 + \cdots + \theta_p\, x_p)$$

$$f\left(X_1 = x_1, \cdots, X_p = x_p\right) = \ln\frac{\pi}{1-\pi} = \theta_0 + \theta_1 x_1 + \cdots + \theta_p\, x_p$$

$$f(X) = \ln\frac{\pi}{1-\pi} = \theta_0 + \theta_1 X_1 + \cdots + \theta_p\, X_p$$



High-order Regression



- High-order polynomial regression model

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \cdots + \theta_m X^m + \epsilon$$
$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \cdots + \theta_m x_i^m + \epsilon_i,\ i = 1,\ \cdots,\ n.$$

- High-order multivariate regression model

$$Y = \theta_0 + \theta_1 X_1 + \cdots + \theta_k X_k + \cdots + \theta_{k(\pi)} X_{\pi_1} \cdots X_{\pi_j} \cdots + \cdots + \theta_p X_{\mu(m)}^m + \epsilon$$
$$y_i = \theta_0 + \theta_1 x_{i1} + \cdots + \theta_k x_{ik} + \cdots + \theta_{k(\pi)} x_{i\pi_1} \cdots x_{i\pi_j} \cdots + \cdots + \theta_M x_{ip}^m + \epsilon_i$$

- Matrix-vector form

Let $\theta = [\theta_0\ \ \theta_1\ \ \cdots\ \theta_p]^T,\ \ \phi_i = [1\ \ \phi_{i1} \cdots \phi_{ip}]^T$

$y = [y_1\ \ y_2\ \ \cdots\ \ y_n]^T,\ \ \ \epsilon = [\epsilon_1\ \ \epsilon_2\ \ \cdots\ \ \epsilon_n]^T$

Then $y_i = \phi_i^T \theta + \epsilon_i,\ i = 1,\ \cdots,\ n.$

$y = \Phi\theta + \epsilon,\ \ \Phi = [\phi_1\ \phi_2\ \ \cdots\ \ \phi_n]^T$

$$\Phi = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^m \\ 1 & x_2 & x_2^2 & \cdots & x_2^m \\ 1 & x_3 & x_3^2 & \cdots & x_3^m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^m \end{bmatrix}$$

Basis-function Regression

- Matrix-vector form of General Regression

Let $\theta = [\theta_0 \;\; \theta_1 \;\; \cdots \; \theta_p]^T, \quad \phi_i = [1 \;\; \phi_{i1} \cdots \phi_{ip}]^T$

$\boldsymbol{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^T, \quad \boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_n]^T$

Then $y_i = \phi_i^T \theta + \epsilon_i, \; i = 1, \; \cdots, \; n.$

$\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \Phi = [\phi_1 \, \phi_2 \;\; \cdots \;\; \phi_n]^T$

- Basis for General Regression
  - sin, cos basis: $\phi_{im} = \sin \omega_m x_i$ or $\cos \omega_m x_i$
  - radial basis: $\phi_{im} = \exp \frac{-\|x_i - \mu_m\|^2}{\sigma_m^2}$
  - sigmoid basis: $\phi_{im} = \frac{1}{1+\exp(-w_m^T x_i - b_m)}$ or $\frac{\exp(w_m^T x_i + b_m)}{1+\exp(w_m^T x_i + b_m)}$

Logistic Regression

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I})$$

MLE:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{\|\epsilon\|^2}{2\sigma^2})$$

LSE:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

Solution:

by $\nabla_\theta S(\theta) = 0$ at $\hat{\theta}$.

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

Solution:

$$\nabla_\theta S(\theta) = 0 \text{ at } \hat{\theta}$$

$$\nabla_\theta (\boldsymbol{y} - \Phi\theta)^T (\boldsymbol{y} - \Phi\theta) = 0 \text{ at } \hat{\theta}$$

$$2\Phi^T \left(\boldsymbol{y} - \Phi\hat{\theta}\right) = 0$$

$$\Phi^T \boldsymbol{y} - \Phi^T\Phi\hat{\theta} = 0$$

$$\hat{\theta} = (\Phi^T\Phi)^{-1} \Phi^T \boldsymbol{y}$$

Interim Summary

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \Phi\theta\|^2 \cong S(\theta)$$

Solution:

$$\nabla_\theta S(\theta) = 0 \text{ at } \hat{\theta}$$

$$\nabla_\theta (\boldsymbol{y} - \Phi\theta)^T (\boldsymbol{y} - \Phi\theta) = 0 \text{ at } \hat{\theta}$$

$$2\Phi^T \left(\boldsymbol{y} - \Phi\hat{\theta}\right) = 0$$

$$\Phi^T \boldsymbol{y} - \Phi^T\Phi\hat{\theta} = 0$$

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^T \boldsymbol{y}$$

Parameter Estimation in Matrix form

- Least Squares Estimation

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\,\Phi^T\,\boldsymbol{y} \leftarrow \boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon}\sim N(0, \sigma^2\mathbf{I})$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_k^T \end{bmatrix} \theta + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_k \end{bmatrix}, \; \Phi_k = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1p} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \phi_{k1} & \phi_{k2} & \cdots & \phi_{kp} \end{bmatrix}$$

$$y_i = \phi_i^T\theta + \epsilon_i$$

$$y_i = \theta_0 + \theta_1\phi_{i1} + \theta_2\phi_{i2} + \cdots + \theta_p\phi_{ip} + \epsilon_i,$$

$$i = 1, \cdots, k, \cdots, n, \cdots$$

- Observation Matrix

$$\Phi_k = [\phi_1\ \phi_2\ \cdots\ \phi_k]^T \rightarrow \underbrace{\Phi_k^T}_{\boxed{(p+1)\times k}}\,\Phi_k = [\phi_1\ \phi_2\ \cdots\ \phi_k]\begin{bmatrix} \phi_1^T \\ \phi_2^T \\ \vdots \\ \phi_k^T \end{bmatrix} = \sum_{i=1}^{k} \phi_i\,\phi_i^T$$

$$\boldsymbol{y}_k = [y_1\ y_2\ \cdots\ y_k]^T$$

- Recursive Least Squares

$$\hat{\theta}_k = (\Phi_k^T\Phi_k)^{-1}\Phi_k^T\boldsymbol{y}_k \rightarrow \hat{\theta}_{k+1} = (\Phi_k^T\Phi_k + \phi_{k+1}\phi_{k+1}^T)^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}$$

Parameter Estimation in Matrix form

- Matrix Inversion Lemma

$$(A+BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1}+CA^{-1}B)^{-1}C\,A^{-1}$$

Sherman-Morrison formula: $(A+uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$

- Recursive Least Squares

$$\hat{\theta}_{k+1} = (\Phi_k^T\Phi_k + \phi_{k+1}\phi_{k+1}^T)^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}$$

define $P_k \cong (\Phi_k^T\Phi_k)^{-1}$,

$$\hat{\theta}_{k+1} = (P_k^{-1} + \phi_{k+1}\phi_{k+1}^T)^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}$$

$$= \left(P_k - \frac{P_k\phi_{k+1}\phi_{k+1}^TP_k}{1+\phi_{k+1}^TP_k\phi_{k+1}}\right)\Phi_{k+1}^T\boldsymbol{y}_{k+1},$$ (don't need inverse)

define $G_k \cong \frac{P_k\phi_{k+1}}{1+\phi_{k+1}^TP_k\phi_{k+1}} \quad \Longrightarrow \quad P_{k+1} = P_k - \frac{P_k\phi_{k+1}\phi_{k+1}^TP_k}{1+\phi_{k+1}^TP_k\phi_{k+1}} = P_k - G_k\phi_{k+1}^TP_k$

Parameter Estimation in Matrix form

- Recursive Least Squares (cont.)

$$\Phi_k = [\phi_1\ \phi_2\ \cdots\ \phi_k]^T$$
$$\mathbf{y}_k = [y_1\ y_2\ \cdots\ y_k]^T$$

$$P_k \cong (\Phi_k^T \Phi_k)^{-1}$$

$$G_k \cong \frac{P_k \phi_{k+1}}{1+\phi_{k+1}^T P_k \phi_{k+1}}$$

$$P_{k+1} = P_k - G_k \phi_{k+1}^T P_k$$

$$\begin{aligned}
\hat{\theta}_{k+1} &= (P_k - G_k \phi_{k+1}^T P_k)[\Phi_k^T \quad \phi_{k+1}] \begin{bmatrix} \mathbf{y}_k \\ y_{k+1} \end{bmatrix} \\
&= (P_k - G_k \phi_{k+1}^T P_k)(\Phi_k^T \mathbf{y}_k + \phi_{k+1} y_{k+1}) \\
&= (I - G_k \phi_{k+1}^T)(P_k \Phi_k^T \mathbf{y}_k + P_k \phi_{k+1} y_{k+1}) \\
&= (I - G_k \phi_{k+1}^T)(\hat{\theta}_k + P_k \phi_{k+1} y_{k+1}) \\
&= \hat{\theta}_k - G_k \phi_{k+1}^T \hat{\theta}_k + P_k \phi_{k+1} y_{k+1} - G_k \phi_{k+1}^T P_k \phi_{k+1} y_{k+1} \\
&= \hat{\theta}_k - G_k \phi_{k+1}^T \hat{\theta}_k + G_k y_{k+1} + G_k \phi_{k+1}^T P_k \phi_{k+1} y_{k+1} - G_k \phi_{k+1}^T P_k \phi_{k+1} y_{k+1}
\end{aligned}$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k + G_k (y_{k+1} - \phi_{k+1}^T \hat{\theta}_k),\ P_0 = \alpha \mathbf{I},\ \alpha \gg 1.$$

$$\hat{\theta} = \hat{\theta}_n \qquad G_k \cong \frac{P_k \phi_{k+1}}{1+\phi_{k+1}^T P_k \phi_{k+1}} \qquad P_{k+1} = P_k - G_k \phi_{k+1}^T P_k$$

Parameter Estimation in Matrix form

- Weighted Recursive Least Squares

$$\hat{\theta}_{k+1} = (\lambda\Phi_k^T\Phi_k + \phi_{k+1}\phi_{k+1}^T)^{-1}\Phi_{k+1}^T\boldsymbol{y}_{k+1}\,,\ 0 < \lambda < 1$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k + G_k(y_{k+1} - \phi_{k+1}^T\hat{\theta}_k),\ P_0 = \alpha\mathbf{I},\ \alpha \gg 1$$

$$\hat{\theta} = \hat{\theta}_n$$

$$G_k \cong \frac{\lambda^{-1}P_k\phi_{k+1}}{1+\lambda^{-1}\phi_{k+1}^T P_k\phi_{k+1}}$$

$$P_{k+1} = \lambda^{-1}P_k - \lambda^{-1}G_k\phi_{k+1}^T P_k$$

$$\Phi_k = [\phi_1\ \phi_2\ \cdots\ \phi_k]^T$$
$$\boldsymbol{y}_k = [y_1\ y_2\ \cdots\ y_k]^T$$

$$P_k \cong (\Phi_k^T\Phi_k)^{-1}$$

$$\lambda^{-1}P_k \cong (\lambda\Phi_k^T\Phi_k)^{-1}$$

Quality of Fit in Matrix form

- Regression model in matrix form

$$\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}$$

- Estimated parameter

$$\hat{\theta} = (\Phi^T\Phi)^{-1}\,\Phi^T\,\boldsymbol{y} = \theta + (\Phi^T\Phi)^{-1}\,\Phi^T\boldsymbol{\epsilon} \text{ (unbiased estimate)}$$

- Confidence Interval

$$E\left(\hat{\theta}\right) = \theta,$$

$$E\left((\theta - \hat{\theta})^T(\theta - \hat{\theta})\right) = E\boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}(\Phi^T\Phi)^{-1}\,\Phi^T\boldsymbol{\epsilon} = E\;Tr(\boldsymbol{\epsilon}^T\Phi(\Phi^T\Phi)^{-1}(\Phi^T\Phi)^{-1}\,\Phi^T\boldsymbol{\epsilon})$$

$$= E\;Tr((\Phi^T\Phi)^{-1}\cancel{(\Phi^T\Phi)^{-1}}\,\cancel{\Phi^T}\Phi\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}) = Tr\left((\Phi^T\Phi)^{-1}\right)\sigma^2 \rightarrow \hat{\theta} = \theta \pm \alpha\sigma$$

- Prediction

$$\hat{\boldsymbol{y}} = \Phi\hat{\theta} = \Phi\theta + \Phi(\Phi^T\Phi)^{-1}\,\Phi^T\boldsymbol{\epsilon} = \Phi\theta + \mathbb{H}\boldsymbol{\epsilon},$$

where $\mathbb{H}$ is symmetric and idempotent ($\mathbb{H}^2 = \mathbb{H}$), $\mathbb{H}\Phi = \Phi$.

$$\mathbb{H}\hat{\boldsymbol{y}} = \mathbb{H}\Phi\theta + \mathbb{H}\boldsymbol{\epsilon} = \Phi\theta + \mathbb{H}\boldsymbol{\epsilon} = \hat{\boldsymbol{y}}$$

- Residual vector : $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbb{H})\boldsymbol{\epsilon}$

Quality of Fit in Matrix form

- Residual vector

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\mathbf{I} - \mathbb{H})\boldsymbol{\epsilon}$$

$$E(\boldsymbol{e}^T\boldsymbol{e}) = E(\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbb{H})(\mathbf{I} - \mathbb{H})\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}^T(\mathbf{I} - \mathbb{H})\boldsymbol{\epsilon})$$
$$= Tr(\mathbf{I} - \mathbb{H})E(\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}) = Tr(\mathbf{I} - \mathbb{H})n\sigma^2 = (n-p-1)n\sigma^2$$

here

$$Tr(\mathbf{I} - \mathbb{H}) = Tr(\mathbf{I}) - Tr(\mathbb{H}) = n - Tr(\Phi(\Phi^T\Phi)^{-1}\,\Phi^T)$$
$$= n - Tr\left((\Phi^T\Phi)^{-1}\,\Phi^T\Phi\right) = n - (p+1), p+1{:}\,\#\ of\ parameters$$

$$\boxed{(p+1)\times n \cdot \mathrm{n}\times\ (p+1)}$$

hence

$$E(\boldsymbol{e}^T\boldsymbol{e}/(n-p-1)) = n\sigma^2 \rightarrow \frac{\boldsymbol{e}^T\boldsymbol{e}}{n-p-1}$$: unbiased estimate of $n\sigma^2$

- Coefficient of Determination

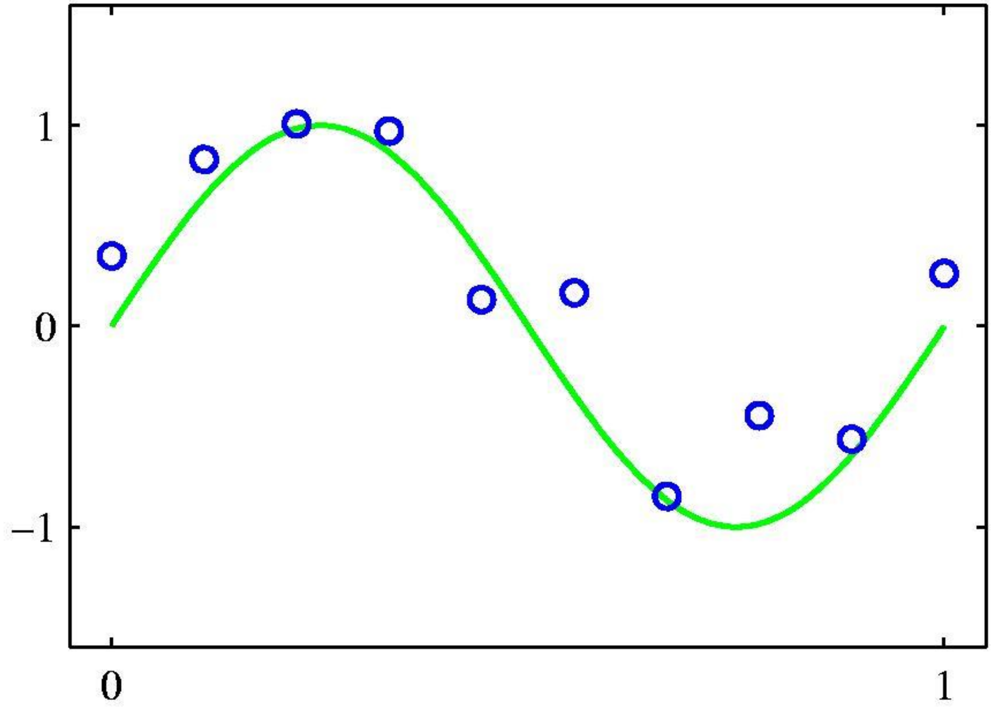
$$R^2 = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}}{(\boldsymbol{y} - \bar{y}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1})},\ R_a^2 = 1 - \frac{\boldsymbol{e}^T\boldsymbol{e}/(n-p-1)}{(\boldsymbol{y} - \bar{y}\mathbf{1})^T(\boldsymbol{y} - \bar{y}\mathbf{1})/(n-1)}$$

PARTIAL LEAST SQUARES REGRESSION

JIN YOUNG CHOI

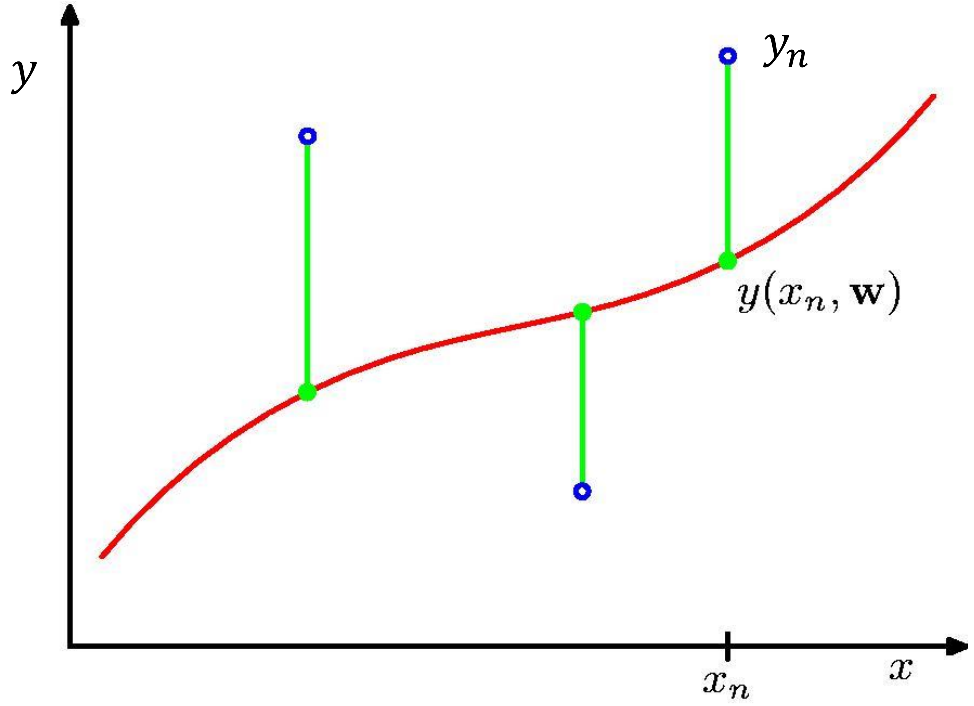
ECE, SEOUL NATIONAL UNIVERSITY

Overfitting and Underfitting

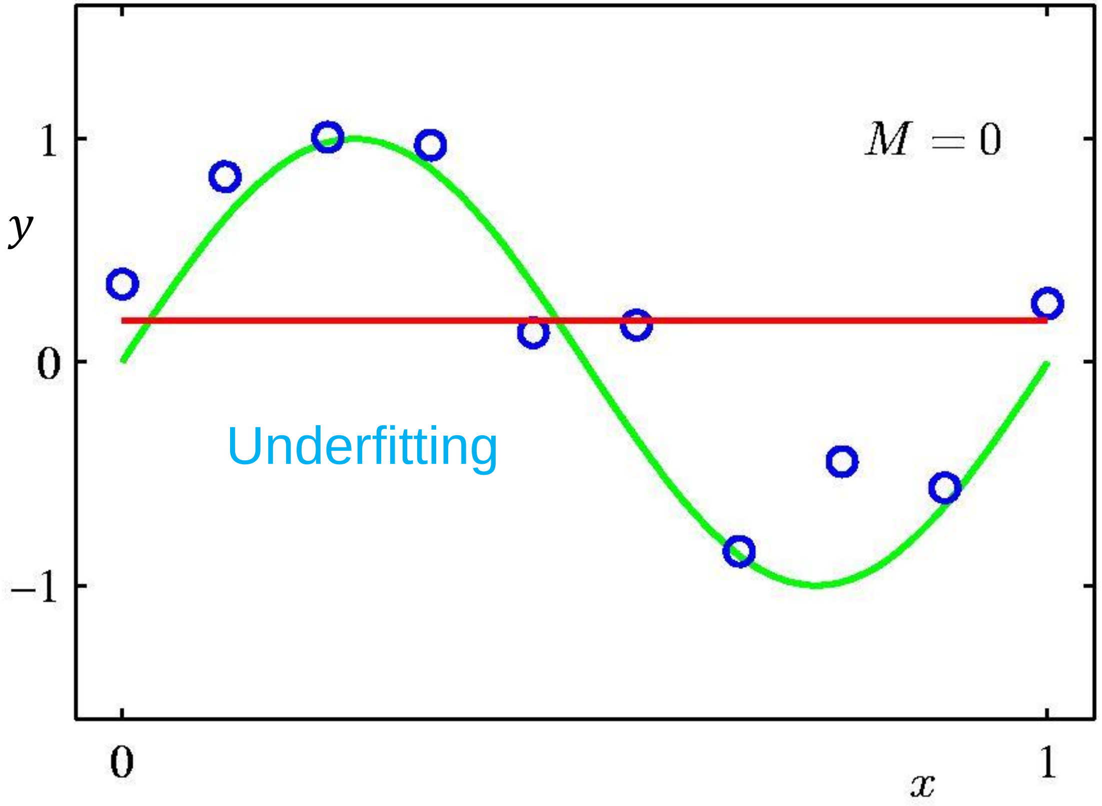


$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \cdots + \theta_M X^M + \epsilon$$

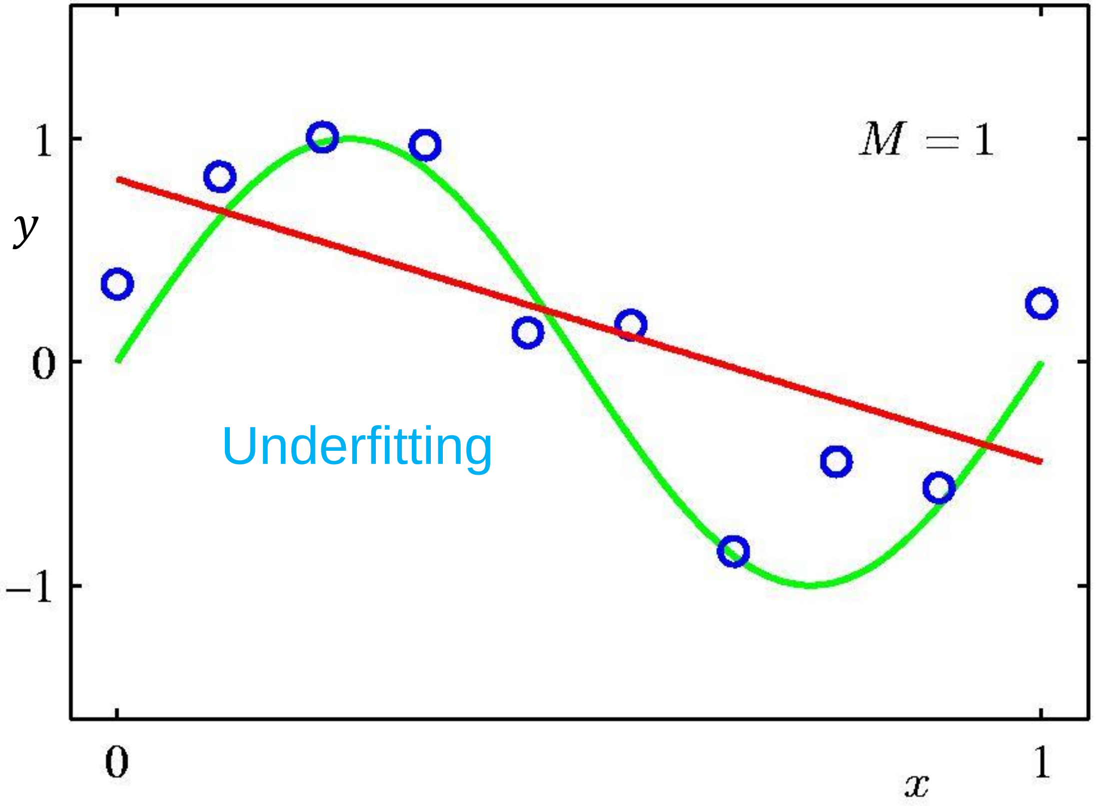
Sum-of-Squares Error Function



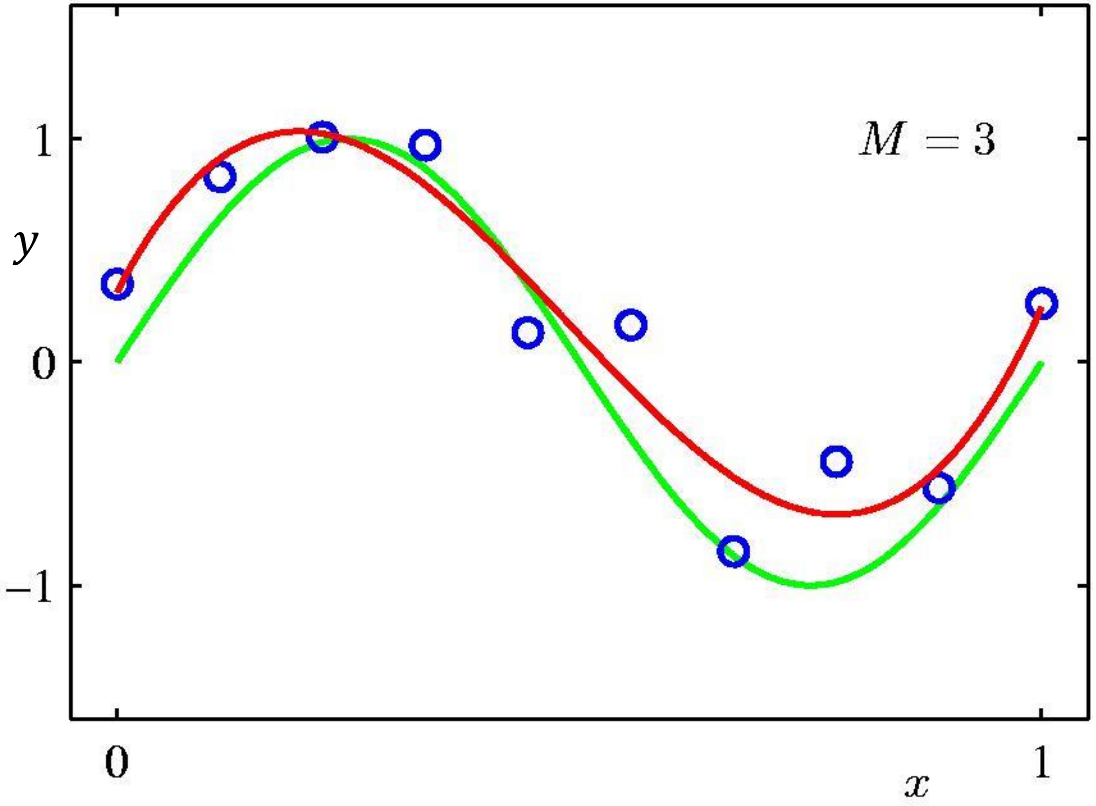
0^{th} Order Polynomial



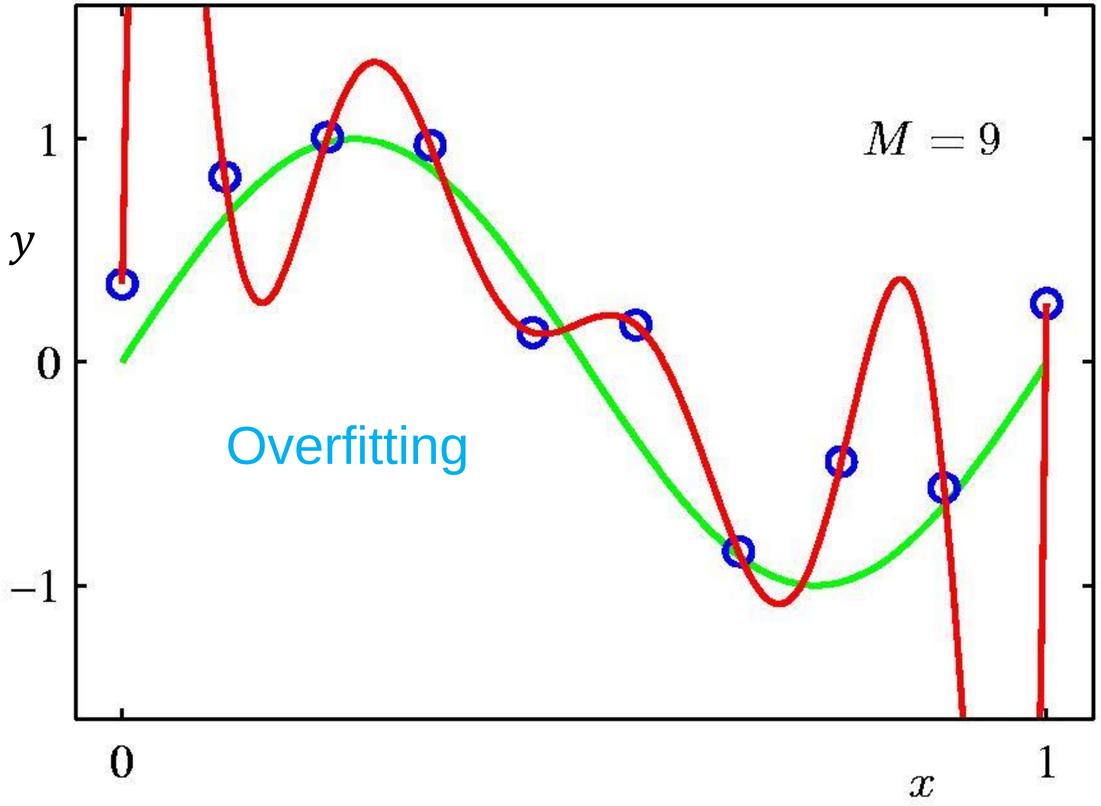
1st Order Polynomial



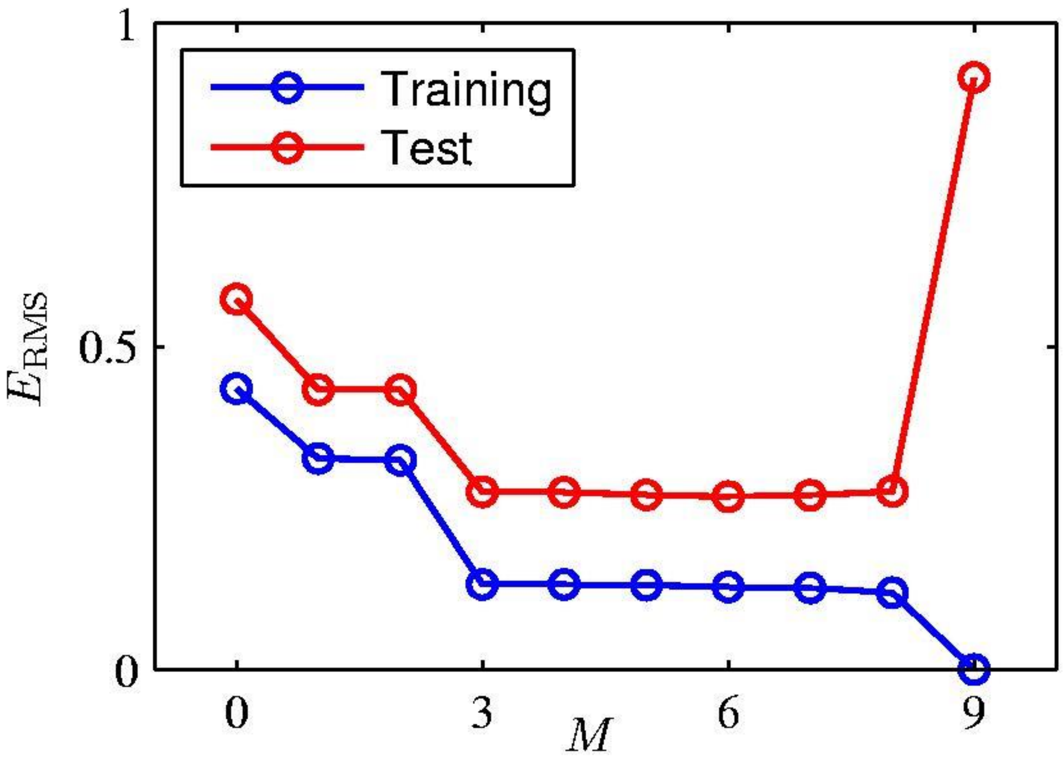
3rd Order Polynomial



9th Order Polynomial



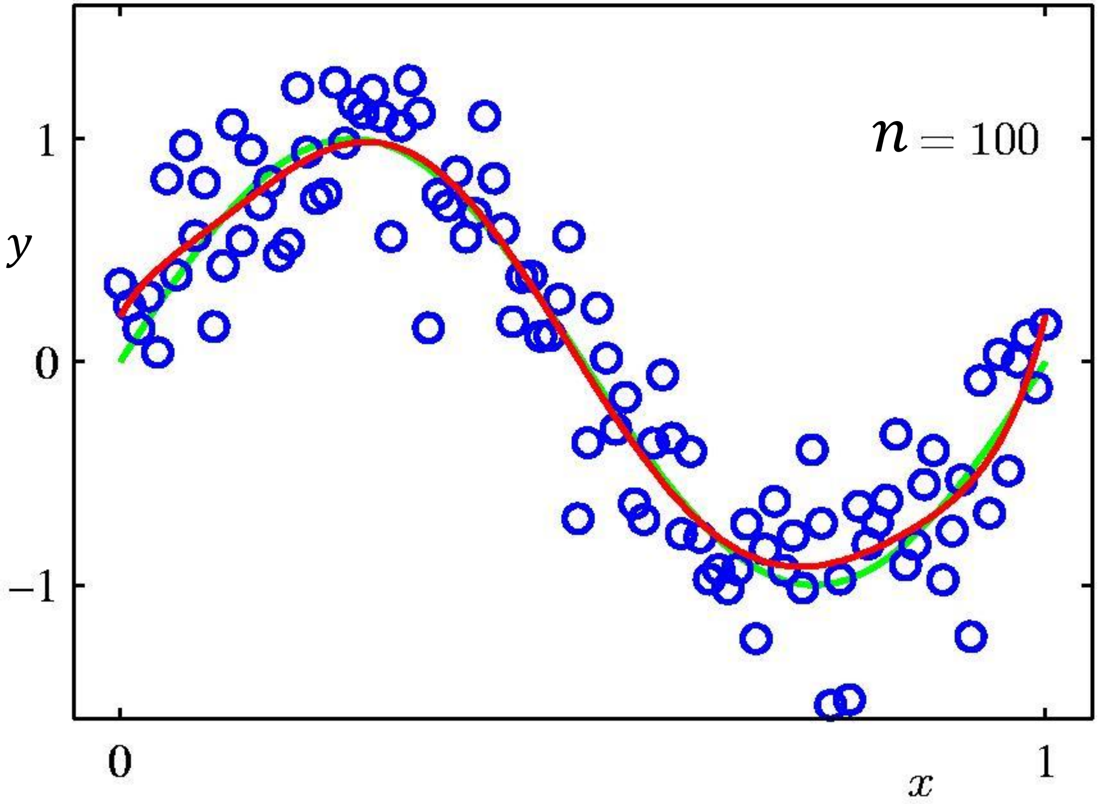
Over-fitting



Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{E(\theta^\star)/n}$

Data Set Size:

9th Order Polynomial



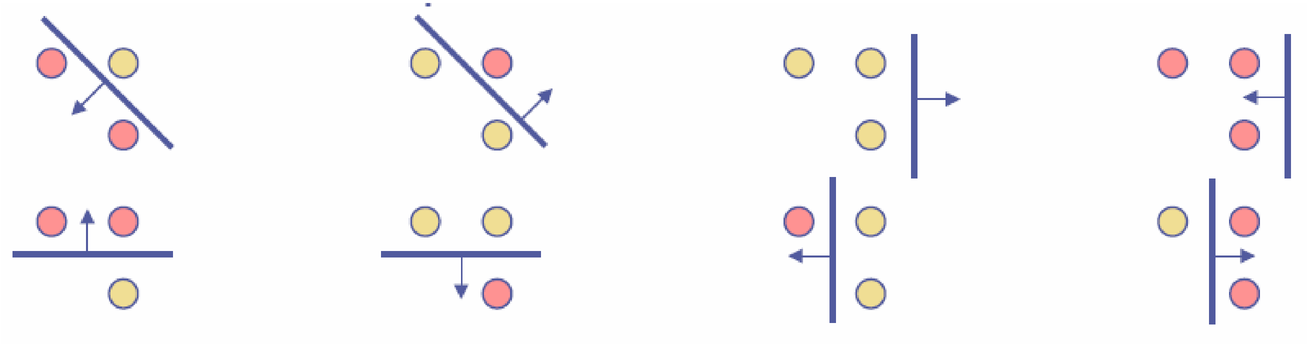
Model Complexity

- VC(Vapnik-Chervonenkis)-dimension:

  Maximum number of points that can be labeled in all possible way

- VC dimension of linear classifiers in *N* dimensions

  is h=N+1 (= #of weights, n_w), cf.) MLP: O($n_w{}^2$)



- Measure of Complexity of a classifier
- **Minimizing VC dim. == Minimizing Complexity**

Bias and Variance in Parameter Estimation

- Mean Squared Error(MSE) decomposition

$$MSE(\hat{\theta}) = E\left((\hat{\theta}-\theta)^2\right)$$

$$= E\left(\left(\hat{\theta}-E(\hat{\theta})+E(\hat{\theta})-\theta\right)^2\right)$$

$$= E\left(\left(\hat{\theta}-E(\hat{\theta})\right)^2 + 2\left(\hat{\theta}-E(\hat{\theta})\right)\left(E(\hat{\theta})-\theta\right) + \left(E(\hat{\theta})-\theta\right)^2\right)$$

$$= E\left(\hat{\theta}-E(\hat{\theta})\right)^2 + 2\,\boxed{\overbrace{E\left(\hat{\theta}-E(\hat{\theta})\right)}^{E(\hat{\theta})-E(\hat{\theta})=0}}\left(E(\hat{\theta})-\theta\right) + \left(E(\hat{\theta})-\theta\right)^2$$

$$= E\left(\hat{\theta}-E(\hat{\theta})\right)^2 + \left(E(\hat{\theta})-\theta\right)^2$$

$$= Var(\hat{\theta}) + Bias(\hat{\theta},\theta)^2$$

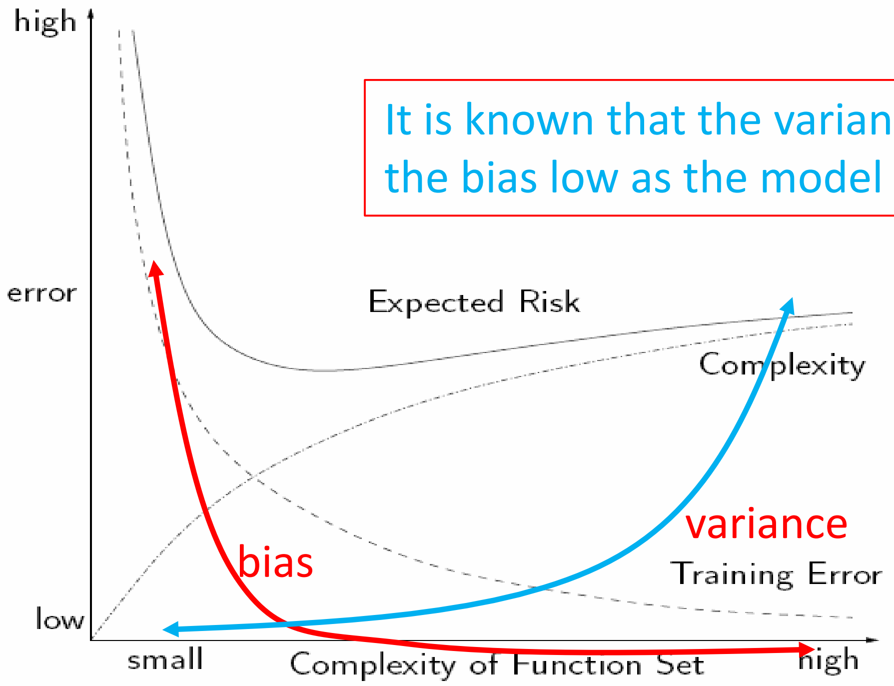
overfitting underfitting

Bias and Variance in Parameter Estimation

- Mean Squared Error(MSE) decomposition

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

$$= Var(\hat{\theta}) + Bias(\hat{\theta}, \theta)^2$$



It is known that the variance becomes high and the bias low as the model complexity increases[†].

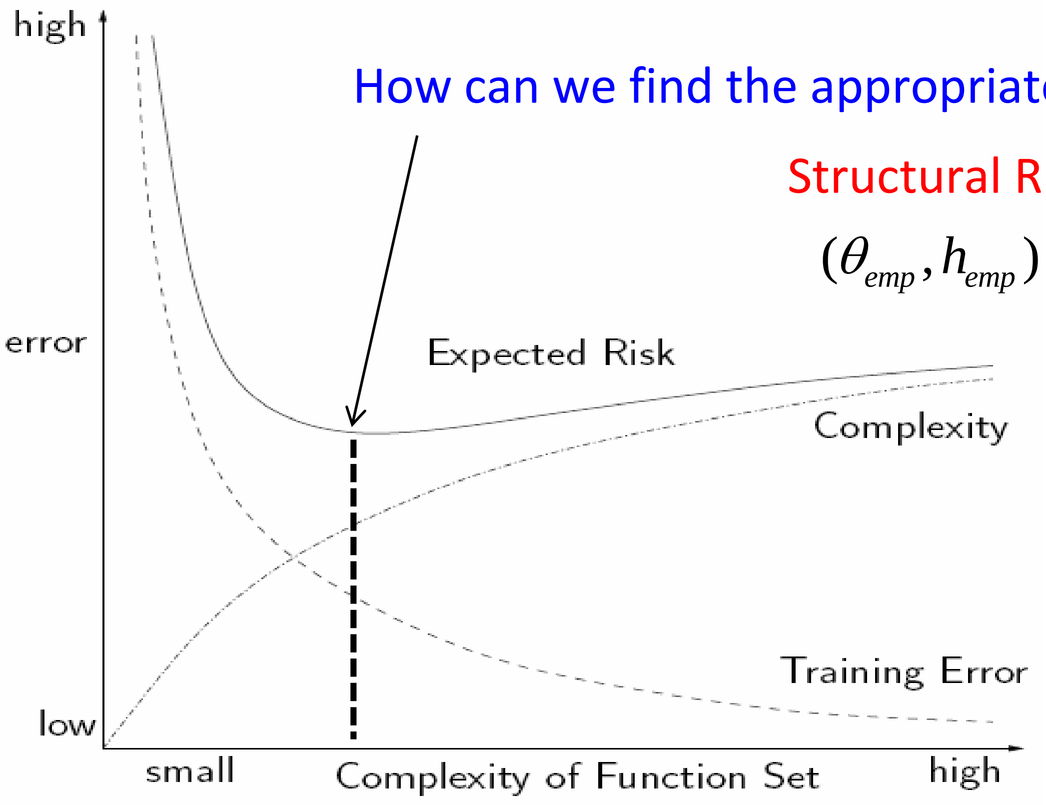
underfitting

overfitting

[†] https://datacadamia.com/data_mining/bias_trade-off

Structural Risk Minimization

- For fixed training samples n



Structural Risk Minimization

$$(\theta_{emp}, h_{emp}) = \arg\min_{\theta,h} R_{emp}(\theta, h)$$

$$\Downarrow$$

Pruning (F-test, PCA)

Regularization
- ridge regression
- Lasso regression

SVM

Partial Least Squares

- Matrix-vector form for General Regression (Revisit)

  Let $\theta = [\theta_0 \;\; \theta_1 \;\; \cdots \; \theta_M]^T, \quad \phi_i = [1 \;\; \phi_{i1} \cdots \phi_{iM}]^T$

  $\boldsymbol{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_n]^T, \quad \boldsymbol{\epsilon} = [\epsilon_1 \;\; \epsilon_2 \;\; \cdots \;\; \epsilon_n]^T$

  Then $y_i = \phi_i^T \theta + \epsilon_i, \;\; i = 1, \; \cdots, \; n.$

  $\boldsymbol{y} = \Phi\theta + \boldsymbol{\epsilon}, \quad \Phi = [\phi_1 \, \phi_2 \;\; \cdots \;\; \phi_n]^T$

- Matrix-vector form for Multivariate Regression with no-intercept

  $y_i = \mathbf{x}_i^T \theta + \epsilon_i, \;\; i = 1, \; \cdots, \; n$

  $\boldsymbol{y} = \mathbf{X}\theta + \boldsymbol{\epsilon}, \;\; \mathbf{X} = [\mathbf{x}_1 \, \mathbf{x}_2 \;\; \cdots \;\; \mathbf{x}_n]^T$

  $\boxed{\begin{array}{l} \mathbf{x}_i = [x_{i1} \cdots x_{ip}]^T, \; \theta = [\theta_1 \;\; \cdots \; \theta_p]^T \\ \mathbf{x}_i = \mathbf{x}_i^o - \mu, \;\; \mu = 1/n \sum_i \mathbf{x}_i^o \end{array}}$

- Goal: reduce the input & parameter dimension: $p > q$

  $\mathbf{x}_i = [x_{i1} \cdots x_{ip}]^T, \; \theta = [\theta_1 \;\; \cdots \; \theta_p]^T \longrightarrow \mathbf{z}_i = [z_{i1} \cdots z_{iq}]^T, \; \theta = [\theta_1 \;\; \cdots \; \theta_q]^T$

Principal Component Regression

$\boxed{\mathbf{a}_k = E^T(\mathbf{x}_k - \mathbf{m})}$

- Principal Component Analysis for $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$

$$\mathbf{S} = \sum_{i=1}^{n} \mathbf{x}_i\, \mathbf{x}_i{}^T = \mathbf{X}\mathbf{X}^T,\ \ \mathbf{S}\mathbf{u}_k = \lambda_k \mathbf{u}_k, \lambda_1 > \lambda_2 \cdots > \lambda_p$$

$$\boxed{\mathbf{cov}(\mathbf{X}, \mathbf{X}) = \frac{1}{n}\mathbf{X}\,\mathbf{X}^T}$$

- Reduced dim. vector ($q < p$ dim.)

$$\underbrace{\boldsymbol{z}_i}_{q\times 1} = \bar{\mathbf{U}}^T \mathbf{x}_i,\ \ \underbrace{\bar{\mathbf{U}}}_{p\times q} = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_q]$$

Orthonormal eigenvectors $\{\mathbf{u}_i\}$ ← **S** is symmetric

$\mathrm{U}^T = \mathrm{U}^{-1}$

$$\mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ \cdots\ \mathbf{z}_n] = \bar{\mathbf{U}}^T\ [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$$

$$\mathbf{Z} = \bar{\mathbf{U}}^T\mathbf{X} \rightarrow \mathbf{Z}^T = \mathbf{X}^T\bar{\mathbf{U}},\quad \boldsymbol{y} = \mathbf{X}^T\theta + \boldsymbol{\epsilon} \approx \boldsymbol{y} = \mathbf{Z}^T\bar{\mathbf{U}}^T\theta + \boldsymbol{\epsilon} = \boldsymbol{y} = \mathbf{Z}^T\vartheta + \boldsymbol{\epsilon}$$

- Applying LS algorithm to $\boldsymbol{y} = \mathbf{Z}^T\vartheta + \boldsymbol{\epsilon}$

$$\hat{\vartheta} = \underset{\vartheta}{\mathrm{argmin}}\|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \mathbf{Z}^T\vartheta\|^2 \rightarrow \boxed{\hat{\vartheta} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\boldsymbol{y} \rightarrow \hat{y} = \boldsymbol{z}^T\hat{\vartheta},\ \ \boldsymbol{z} = \bar{\mathbf{U}}^T\mathbf{x}}$$

Partial Least Squares

- Nonlinear Iterative Partial Least Squares (NIPALS) algorithm

$$\mathbf{X}\mathbf{X}^T\mathbf{u} = \lambda\mathbf{u}$$

Let $\mathbf{t} = \mathbf{X}^T\mathbf{u}$

$$\mathbf{u} = \frac{1}{\lambda}\mathbf{X}\mathbf{t}$$

Since $\|\mathbf{u}\| := 1 = \frac{1}{\lambda}\|\mathbf{X}\mathbf{t}\|$

$$\lambda = \|\mathbf{X}\mathbf{t}\|$$

$$\bar{\mathbf{U}} = [\mathbf{u}_1\ \mathbf{u}_2\ \cdots\ \mathbf{u}_q]\ ,\ \boldsymbol{z}_i = \bar{\mathbf{U}}^T\mathbf{x}_i$$

$$\mathbf{Z} = \bar{\mathbf{U}}^T\mathbf{X} \rightarrow \mathbf{Z}^T = \mathbf{X}^T\bar{\mathbf{U}},\quad \mathbf{Z} = [\mathbf{z}_1\ \mathbf{z}_2\ \cdots\ \mathbf{z}_{\boldsymbol{n}}]$$

$\mathbf{t} := \mathbf{x}_j$ for some j

Loop

$$\mathbf{u} = \mathbf{X}\mathbf{t}/\|\mathbf{X}\mathbf{t}\|$$

$$\mathbf{t} = \mathbf{X}^T\mathbf{u}$$

Until $\mathbf{t}$ stop changing

$$\mathbf{X}^T := \mathbf{X}^T - \mathbf{t}\mathbf{u}^{\boldsymbol{T}} = \mathbf{X}^T(\mathrm{I} - \mathbf{u}\mathbf{u}^{\boldsymbol{T}})$$

Repeat the Loop up to a small $\|\mathbf{X}\mathbf{t}\|$

- Applying LS algorithm to $\boldsymbol{y} = \mathbf{Z}^T\vartheta + \boldsymbol{\epsilon}$

$$\hat{\vartheta} = \underset{\vartheta}{\operatorname{argmin}}\|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \mathbf{Z}^T\vartheta\|^2 \rightarrow \hat{\vartheta} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\boldsymbol{y} \rightarrow \hat{y} = \boldsymbol{z}^T\hat{\vartheta},\ \boldsymbol{z} = \bar{\mathbf{U}}^T\mathbf{x}$$

Ridge Regression for Regularization

- l_2 regularization term is added

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\|\theta\|_2^2 \; (= S(\theta))$$

- solution:

$$\nabla_\theta\left((\boldsymbol{y} - \Phi\theta)^T (\boldsymbol{y} - \Phi\theta) + \gamma\theta^T\theta\right) = 0 \text{ at } \hat{\theta}$$

$$2\Phi^T\left(\boldsymbol{y} - \Phi\hat{\theta}\right) + 2\gamma\hat{\theta} = 0$$

$$\boxed{\hat{\theta} = (\Phi^T\Phi - \gamma\mathbf{I})^{-1}\,\Phi^T\,\boldsymbol{y}}$$

$$\hat{\theta}_{k+1} = \hat{\theta}_k + G_k(y_{k+1} - \phi_{k+1}^T\hat{\theta}_k),$$

$$G_k \cong \frac{\lambda^{-1}P_k\phi_{k+1}}{1+\lambda^{-1}\phi_{k+1}^T P_k\phi_{k+1}}$$

$$P_{k+1} = \lambda^{-1}P_k - \lambda^{-1}G_k\phi_{k+1}^T P_k,\; P_0 = -\gamma\mathbf{I}$$

$$P_0 = \alpha\mathbf{I},\; \alpha \gg 1$$

Lasso Regression for Regularization

- LASSO(Least Absolute Shrinkage Selector Operator)
- l_1 regularization term is added

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\|\theta\|_1$$

- solution: l_1 norm is not differentiable→ constrained convex form

  by adding new optimization variables,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\boldsymbol{1}^T\boldsymbol{s}$$

$$\text{subject to} \quad |\theta_i| \le s_i, \quad i = 1, \cdots, n$$

$$\boxed{\begin{aligned} &\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma\boldsymbol{1}^T\boldsymbol{s} \\ &\qquad \text{subject to} \;\; -s_i \le \theta_i \le s_i, \quad i = 1, \cdots, n \end{aligned}}$$

Elastic Regression for Regularization

- Ridge + LASSO

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma_1 \|\theta\|_2^2 + \gamma_2 \|\theta\|_1$$

- solution: l_1 norm is not differentiable→ constrained convex form

  by adding new optimization variables,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma_1 \|\theta\|_2^2 + \gamma_2 \mathbf{1}^T \boldsymbol{s}$$

$$\text{subject to} \quad |\theta_i| \le s_i, \quad i = 1, \cdots, n$$

$$\boxed{\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmin}} \|\boldsymbol{y} - \Phi\theta\|^2 + \gamma_1 \|\theta\|_2^2 + \gamma_2 \mathbf{1}^T \boldsymbol{s} \\ &\text{subject to} \quad -s_i \le \theta_i \le s_i, \quad i = 1, \cdots, n \end{aligned}}$$

Interim Summary

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression

Outline

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression

GAUSSIAN PROCESS REGRESSION

JIN YOUNG CHOI

ECE, SEOUL NATIONAL UNIVERSITY

https://arxiv.org/pdf/2009.10862.pdf
https://github.com/jwangjie/Gaussian-Processes-Regression-Tutorial
http://mlg.eng.cam.ac.uk/tutorials/06/es.pdf
https://www.sciencedirect.com/science/article/abs/pii/S0022249617302158
http://www.gaussianprocess.org/gpml/chapters/RW.pdf

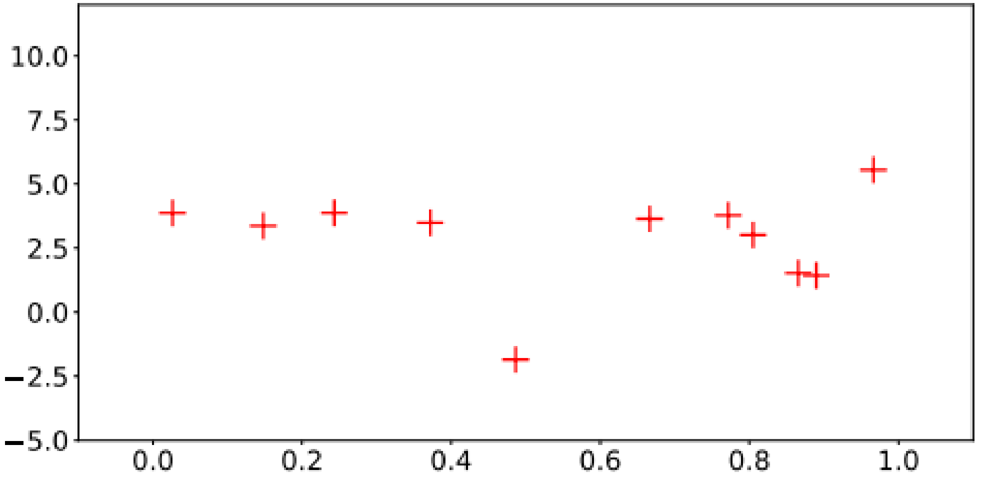
Gaussian Process Regression

- General regression model (single variable)

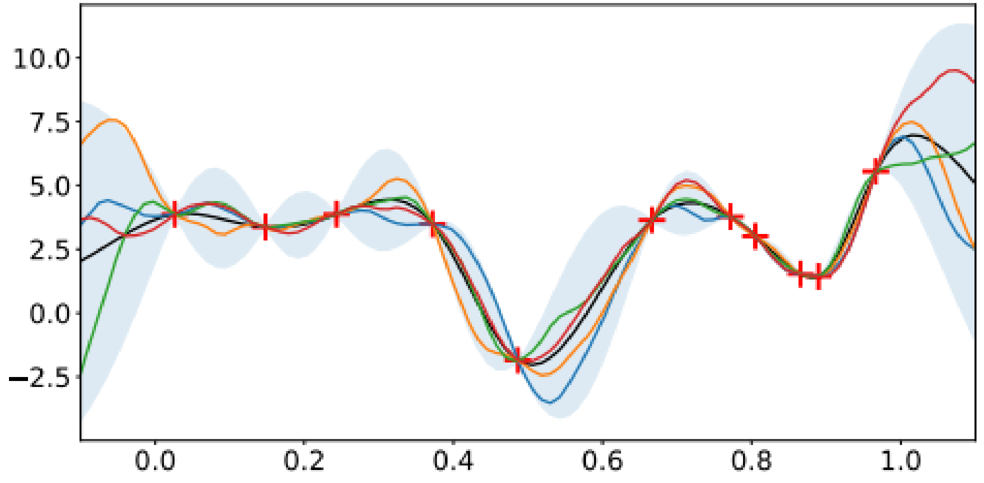
$$y = f(x) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and so x, y are Gaussian random variables.

- Goal : to estimate $f(x)$ with uncertainty from observation data $D = \{(x_i, y_i) | i = 1, \cdots, n\}$
- x_i, y_i are treated as Gaussian random variables.



(a) Data point observations



(b) Five possible regression functions by GPR

Gaussian Process Regression

- General regression model (single variable)

$$y = f(x) + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and so x, y are Gaussian random variables.

- Define

$$\mathbf{x}^{\mathrm{T}} = [x_1 \quad \cdots \quad x_n], \qquad \mathbf{y}^{\mathrm{T}} = [y_1 \quad \cdots \quad y_n], \qquad \mathbf{f} := \mathbf{f}(\mathbf{x}) = [f(x_1) \quad \cdots \quad f(x_n)].$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left[-\frac{1}{2} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \boldsymbol{\mu} \right)^T \Sigma^{-1} \left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \boldsymbol{\mu} \right) \right] := \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \Sigma)$$

- Conditional probability (recall)

$$f_{X|Y}(x|y) = \frac{1}{(2\pi)^{\frac{k}{2}} \sqrt{\det \Sigma_{X|y}}} \exp\left(-\frac{1}{2} (x - \mu_{X|y})^t \Sigma_{X|y}^{-1} (x - \mu_{X|y}) \right),$$

where

$$\mu_{X|y} = A(y - \mu_Y) + \mu_X \text{ and}$$

$$\Sigma_{X|y} = \Sigma_X - AC_{YX}, \text{ where } A\Sigma_Y = \Sigma_{XY}.$$

Gaussian Process Regression

- General regression model (single variable)

$$y = f(x) + \epsilon,$$

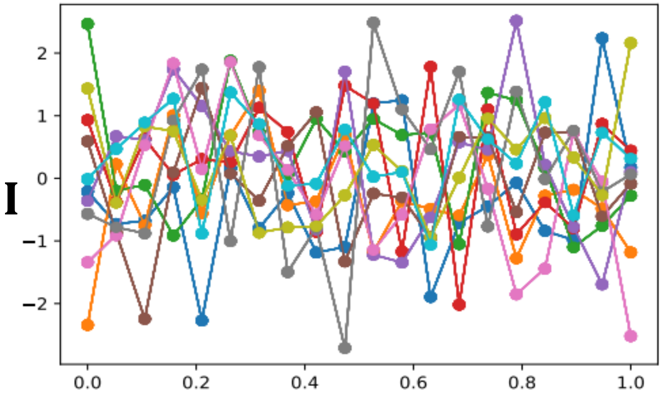
where $\epsilon \sim N(0, \sigma^2)$ and so x, y are Gaussian random variables.

- Define

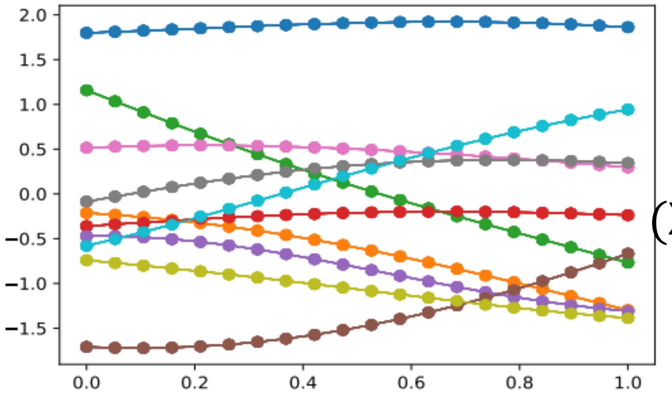
$$\mathbf{x} = [x_1 \quad \cdots \quad x_n], \qquad \mathbf{y} = [y_1 \quad \cdots \quad y_n], \qquad \mathbf{f} := \mathbf{f}(\mathbf{x}) = [f(x_1) \quad \cdots \quad f(x_n)].$$

$$p(\mathbf{x}|\mathbf{y}) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_{\mathbf{x}|\mathbf{y}}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \mu_{\mathbf{x}|\mathbf{y}})^T {\Sigma_{\mathbf{x}|\mathbf{y}}}^{-1}(\mathbf{x} - \mu_{\mathbf{x}|\mathbf{y}})\right] := \boldsymbol{\mathcal{N}}(\mu_{\mathbf{x}|\mathbf{y}}, \Sigma_{\mathbf{x}|\mathbf{y}})$$

$\Sigma_{\mathbf{x}|\mathbf{y}} = \sigma^2 \mathbf{I}$



Independent Gaussian



Jointly Gaussian

$x = f^{-1}(y)$

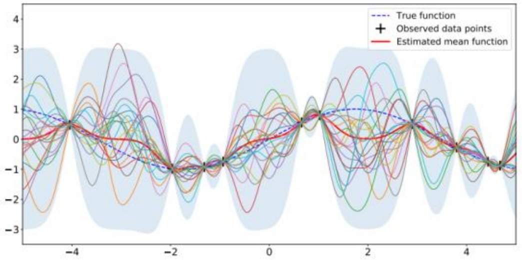
$\Sigma_{\mathbf{x}|\mathbf{y}} \neq \sigma^2 \mathbf{I}$

$(\Sigma_{\mathbf{x}|\mathbf{y}})_{ij} = cov(x_i, x_j)$

$:= exp\left(-\frac{(x_i - x_j)^{\,2}}{2}\right)$

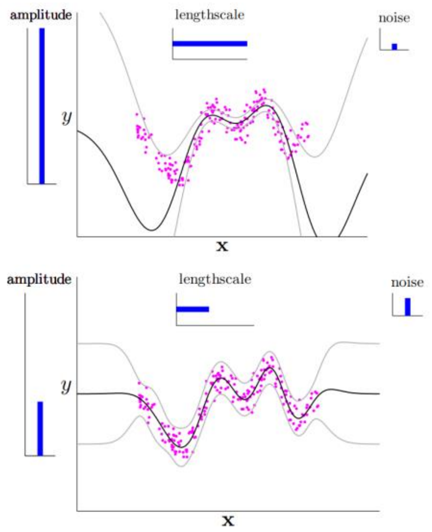
a RBF kernel

Gaussian Process Regression



- Gaussian Processes ($\mathcal{GP}$) for multivariate regression
$$y = f(\mathbf{x}) + \epsilon.$$

- define $\mu_f(\mathbf{x}) := \mathbb{E}(f(\mathbf{x}))$, then we assume $f(\mathbf{x})$ is distributed as a Gaussian process
$$f(\mathbf{x}) \sim \mathcal{GP}\left(\mu_f(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$
where $k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\left(f(\mathbf{x}) - \mu_f(\mathbf{x})\right)\left(f(\mathbf{x}') - \mu_f(\mathbf{x}')\right)\right]$ called the kernel of $\mathcal{GP}$.



- The kernel is based on assumptions such as smoothness, that is, similar $\mathbf{x}, \mathbf{x}'$ yields similar $f(\mathbf{x})$ and $f(\mathbf{x}')$. Thus a popular kernel is
$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \mathbf{exp}\left(-\frac{1}{2\lambda}(\mathbf{x} - \mathbf{x}')^T (\mathbf{x} - \mathbf{x}')\right),$$
where hyperparameters λ and σ_f^2 represents the length-scale and signal (f) variance to control relation between $\mathbf{x}$ and $f(\mathbf{x})$.

Gaussian Process Regression

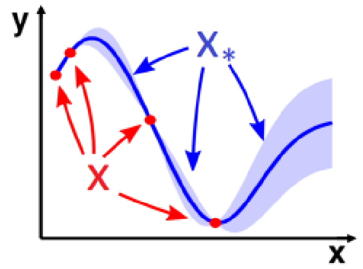
$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda}(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')\right)$$

Modeling of prior sampling function of $\mathcal{GP}$

- Denote $\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n]$, $\mathbf{y}^T = [y_1 \quad \cdots \quad y_n]$, $\mathbf{f}^T := [f(\mathbf{x}_1) \quad \cdots \quad f(\mathbf{x}_n)]$.

  Let $\mathbf{X}_*$ be a matrix containing a new input points $\mathbf{x}_i^*, i = 1, \cdots, n$. Then define the kernel matrix as

$$\mathbf{K}(\mathbf{X}_*,\mathbf{X}_*) = \begin{bmatrix} k(\mathbf{x}_1^*,\mathbf{x}_1^*) & k(\mathbf{x}_1^*,\mathbf{x}_2^*) & \cdots & k(\mathbf{x}_1^*,\mathbf{x}_n^*) \\ k(\mathbf{x}_2^*,\mathbf{x}_1^*) & k(\mathbf{x}_2^*,\mathbf{x}_2^*) & \cdots & k(\mathbf{x}_2^*,\mathbf{x}_n^*) \\ \vdots & \vdots & \ddots & \vdots \\ k(\mathbf{x}_n^*,\mathbf{x}_1^*) & k(\mathbf{x}_n^*,\mathbf{x}_2^*) & \cdots & k(\mathbf{x}_n^*,\mathbf{x}_n^*) \end{bmatrix}$$



- Choosing the prior mean function $\mu_f(\mathbf{x}) = 0$, we can sample values of f at inputs $\mathbf{X}_*$ from $\mathcal{GP}$ as

$$\mathbf{f}_* \sim \mathcal{N}\left(0, \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*)\right)$$

  which is the prior distribution model without observation data $D = \{(x_i, y_i) | i = 1, \cdots, n\}$.

Gaussian Process Regression

Posterior predictions from a $\mathcal{GP}$

- Observations are $D = \{(\mathbf{x}_i, y_i) | i = 1, \cdots, n\} = \{\mathbf{X}, \mathbf{y}\}$, $\mathbf{X} = [\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n]$, $\mathbf{y}^T = [y_1 \quad \cdots \quad y_n]$.
- The predictions for new inputs $\mathbf{X}_*$ by drawing $\mathbf{f}_*$ from the posterior distribution $p(f \mid D)$.

  A joint Gaussian distribution of $\mathbf{y}$ and $\mathbf{f}_*$ Let $\mathbf{X}_*$ follows

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\epsilon^2 \mathbf{I} & \mathbf{K}(\mathbf{X}, \mathbf{X}_*) \\ \mathbf{K}(\mathbf{X}_*, \mathbf{X}) & \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right),$$

$$\boxed{\begin{aligned} y &= f(x) + \epsilon \\ y_* &= f_*(x_*) + \epsilon \end{aligned}}$$

  where σ_ϵ^2 is the assumed noise level of the observations.

- The conditional distribution $p(\mathbf{f}_* | \mathbf{X}, \mathbf{y}, \mathbf{X}_*)$ can be derived to a multivariate normal distribution with mean

$$\mu_{\mathbf{f}_*}(\mathbf{X}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{y}$$

  and variance

$$cov_{\mathbf{f}_*}(\mathbf{X}_*) = \mathbf{K}(\mathbf{X}_*, \mathbf{X}_*) - \mathbf{K}(\mathbf{X}_*, \mathbf{X})[\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_\epsilon^2 \mathbf{I}]^{-1} \mathbf{K}(\mathbf{X}, \mathbf{X}_*)$$

Gaussian Process Regression

Posterior predictions from a $\mathcal{GP}$

- The mean function of the $\mathcal{GP}$ can be given as

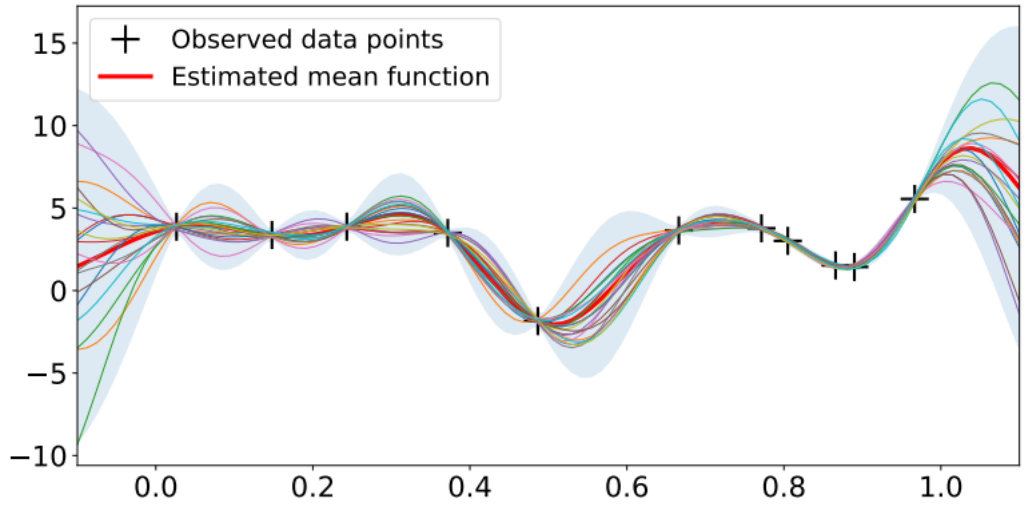
$$\mu_f(\mathbf{x}) = \mathbf{K}(\mathbf{x},\mathbf{X})[\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma_\epsilon^2\mathbf{I}]^{-1}\mathbf{y}$$

and covariance function as

$$cov_f(\mathbf{x},\mathbf{x}') = k(\mathbf{x},\mathbf{x}') - \mathbf{K}(\mathbf{x},\mathbf{X})[\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma_\epsilon^2\mathbf{I}]^{-1}\mathbf{K}(\mathbf{X},\mathbf{x}')$$

$$\mathbf{K}(\mathbf{x},\mathbf{X}) = [k(\mathbf{x},\mathbf{x}_1) \quad k(\mathbf{x},\mathbf{x}_2) \quad \cdots \quad k(\mathbf{x},\mathbf{x}_n)]$$

$$\mathbf{K}(\mathbf{X},\mathbf{x}') = \begin{bmatrix} k(\mathbf{x}_1,\mathbf{x}') \\ k(\mathbf{x}_2,\mathbf{x}') \\ \vdots \\ k(\mathbf{x}_n,\mathbf{x}') \end{bmatrix}$$

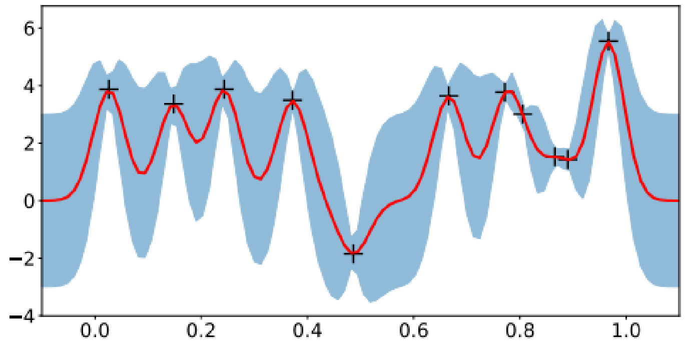


Gaussian Process Regression

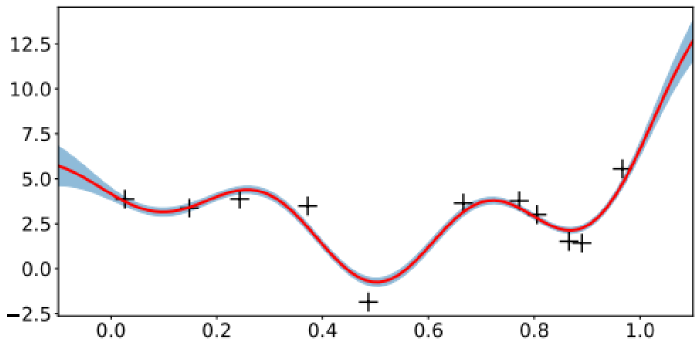
- The effect of the hyperparameters λ and σ_f^2 of the kernel

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \,\mathbf{exp}\left(-\frac{1}{2\lambda}(\mathbf{x}-\mathbf{x}')^T\,(\mathbf{x}-\mathbf{x}')\right) \approx \mathbb{E}\left[\left(f(\mathbf{x})-\mu_f(\mathbf{x})\right)\left(f(\mathbf{x}')-\mu_f(\mathbf{x}')\right)\right],$$

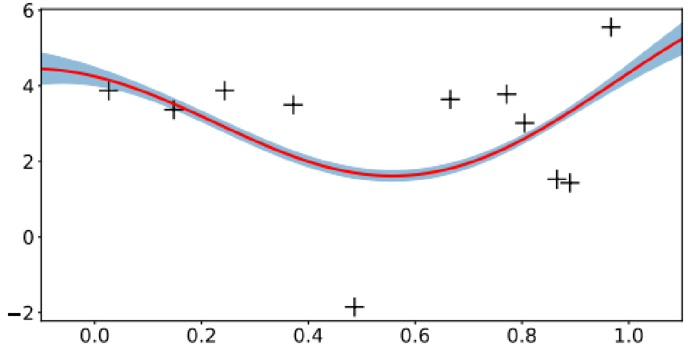
λ : length-scale, σ_f^2 : signal (f) variance to control relation between $\mathbf{x}$ and $f(\mathbf{x})$.



Small λ



Medium λ



Large λ

Gaussian Process Regression

$$p(\mathbf{y}|\mathbf{X}) = \frac{1}{(2\pi)^{d/2}\left|\Sigma_{\mathbf{y}|X}\right|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mu_{\mathbf{y}|X})^T \Sigma_{\mathbf{y}|\mathbf{X}}{}^{-1}(\mathbf{y} - \mu_{\mathbf{y}|X})\right]$$

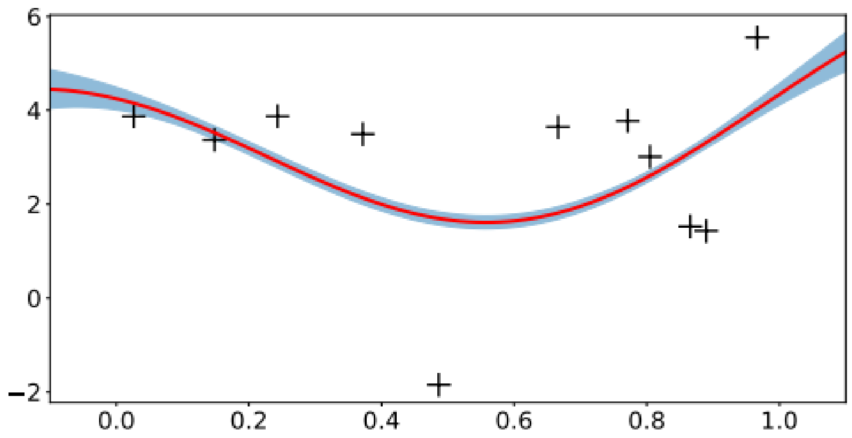
$$\boxed{\begin{bmatrix}\mathbf{y}\\ \mathbf{f}_*\end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma_\epsilon^2\mathbf{I} & \mathbf{K}(\mathbf{X},\mathbf{X}_*)\\ \mathbf{K}(\mathbf{X}_*,\mathbf{X}) & \mathbf{K}(\mathbf{X}_*,\mathbf{X}_*)\end{bmatrix}\right)}$$

$$\boxed{k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda}(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')\right)}$$

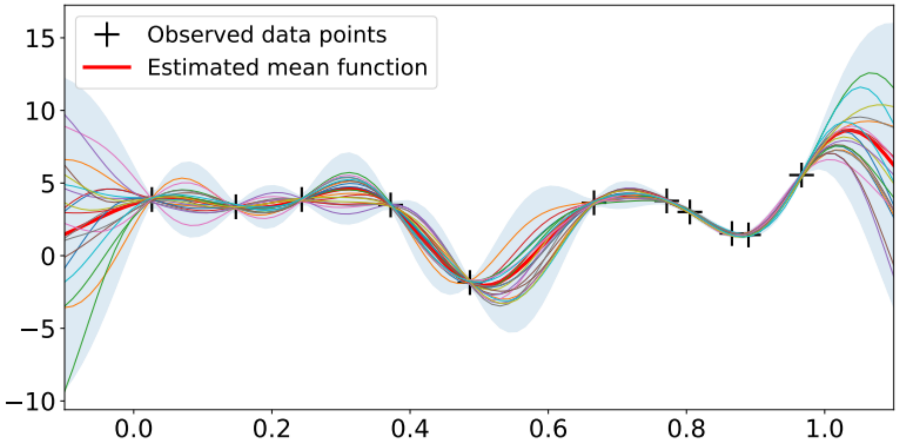
- The optimized hyperparameters λ and σ_f^2

$$\lambda, \sigma_f^2 = \max_{\lambda,\, \sigma_f^2} \log p(\mathbf{y}|\mathbf{X})$$

$$\log p(\mathbf{y}|\mathbf{X}) = -\frac{1}{2}\mathbf{y}^T[\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma_\epsilon^2\mathbf{I}]^{-1}\mathbf{y} - \frac{1}{2}\log\det[\mathbf{K}(\mathbf{X},\mathbf{X}) + \sigma_\epsilon^2\mathbf{I}] - \frac{n}{2}\log 2\pi$$



$\sigma_f = 0.0067$
$\lambda = 0.0967$



Summary

- linear regression
  - simple linear regression
  - multiple linear regression
- nonlinear regression
  - logistic regression
  - high-order regression
  - basis-function regression
- matrix form for regression
  - recursive least squares
- partial least squares
  - over-fitting and underfitting
  - bias/variance
  - principle component regression
  - partial least squares algorithm
  - ridge regression
  - lasso, elastic regression
- Gaussian process regression