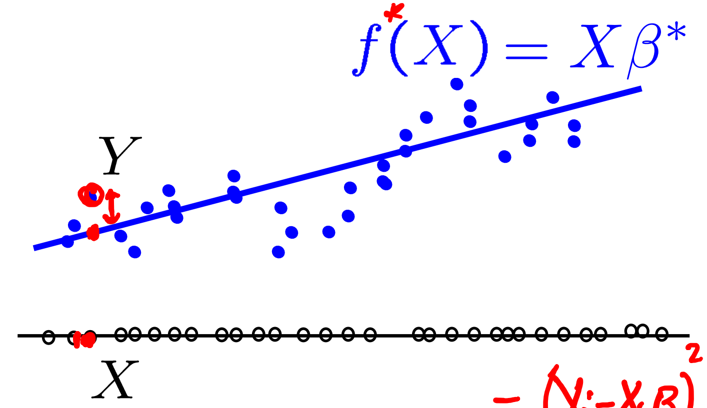
Least Squares and M(C)LE

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(f(X_i)-Y_i\right)^2}_{\text{least squares}}$$

Intuition: Signal plus (zero-mean) Noise model



$$Y = f^*(X) + \epsilon = X\beta^* + \epsilon$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \qquad Y \sim \mathcal{N}(X\beta^*, \sigma^2\mathbf{I})$$

$P(Y|X) =$

$$P(Y_i|X_i) \propto e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}$$

$$\widehat{\beta}_{\text{MLE}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}}$$

(C)

$$= \arg\max_{\beta} \log \prod_{i=1}^{n} p(Y_i|X_i) = \arg\max_{\beta} \sum_{i=1}^{n} -\frac{(Y_i - X_i\beta)^2}{2\sigma^2}$$

$$= \arg\min_{\beta} \sum_{i=1}^{n} (X_i\beta - Y_i)^2 = \widehat{\beta}$$

Least Square Estimate is same as Maximum Conditional Likelihood Estimate under a Gaussian model !

Regularized Least Squares and M(C)AP

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

$\text{Reg} \equiv \text{pen}(\beta) = \|\beta\|_2^2$ or $\|\beta\|_1$

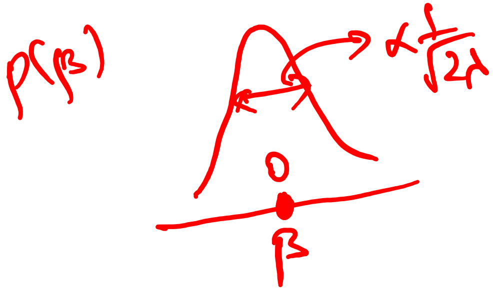
$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

$$= \min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2$$

$\log p(\beta) \propto -\lambda\|\beta\|_2^2$

$p(\beta) \propto e^{-\lambda\|\beta\|_2^2}$

$\sim N\left(0, \frac{1}{\sqrt{2\lambda}}\right)$



$\beta \sim N(0, \sigma^2)$

$e^{-\frac{\|\beta\|^2}{2\sigma^2}} \quad \sigma^2 \equiv \frac{1}{\lambda}$

Regularized Least Squares and M(C)AP

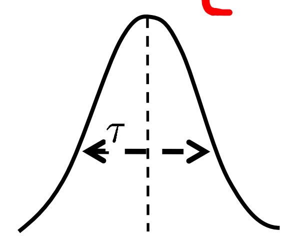
What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n|\beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

$\beta \sim \mathcal{N}(0, \Sigma)$ $\quad p(\beta) \propto e^{-\frac{\beta^T \Sigma^{-1} \beta}{2}}$ $\quad \Sigma = \begin{bmatrix} \frac{1}{\sqrt{2\pi}} & 0 \\ 0 & \frac{1}{\sqrt{2\pi}} \end{bmatrix}$

I) Gaussian Prior

$$\beta \sim \mathcal{N}(0, \tau^2\mathbf{I}) \qquad p(\beta) \propto e^{-\beta^T\beta/2\tau^2}$$



$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda \|\beta\|_2^2$$

Ridge Regression

$\lambda \downarrow$ constant(σ^2, τ^2)

$$\widehat{\beta}_{\text{MAP}} = (\boldsymbol{A}^\top \boldsymbol{A} + \lambda \boldsymbol{I})^{-1} \boldsymbol{A}^\top \boldsymbol{Y}$$

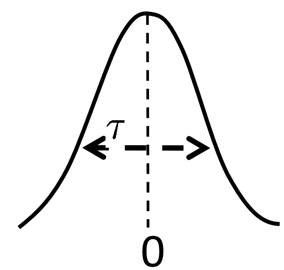
Regularized Least Squares and M(C)AP

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

I) Gaussian Prior

$$\beta \sim \mathcal{N}(0, \tau^2\mathbf{I}) \qquad p(\beta) \propto e^{-\beta^T\beta/2\tau^2}$$



$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \underbrace{\lambda}_{\text{constant}(\sigma^2,\tau^2)} \|\beta\|_2^2$$

Ridge Regression

Prior belief that β is Gaussian with zero-mean biases solution to “small” β

Regularized Least Squares and M(C)AP

Lasso

$$\arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_1$$

What if $(\mathbf{A}^T\mathbf{A})$ is not invertible ?

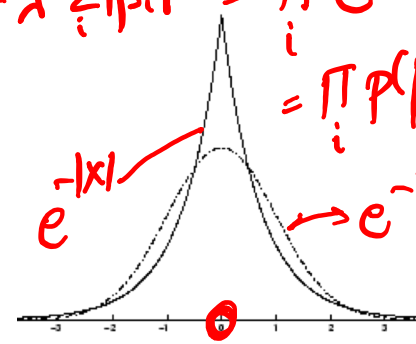
$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n | \beta, \sigma^2, \{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

$$-\log p(\beta) \propto \lambda \sum_i |\beta_i| \Rightarrow p(\beta) \propto e^{-\lambda \sum_i |\beta_i|} = \prod_i e^{-\lambda|\beta_i|} = \prod_i p(\beta_i)$$

II) Laplace Prior

$$\beta_i \overset{iid}{\sim} \text{Laplace}(0, t) \qquad t = \frac{1}{\lambda}$$

$$p(\beta_i) \propto e^{-|\beta_i|/t}$$



$$\widehat{\beta}_{\text{MAP}} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_1$$

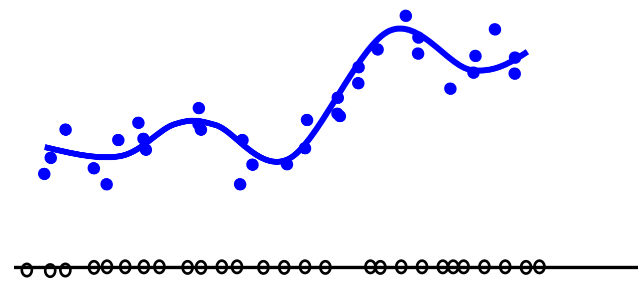
Lasso

$\lambda \rightarrow$ constant(σ^2, t)

Prior belief that β is Laplace with zero-mean biases solution to “sparse” β

Beyond Linear Regression

- Polynomial regression
- Regression with nonlinear features



- Kernelized Ridge Regression (Later)

- Local Kernel Regression (Later)

$(AA^T)^{-1}$

$X = [x(1) \, . \, . \, x(p)]$

$\rightarrow X = [\underbrace{x(1) \, . \, . \, x(p) \; x^2(1) \, - - \, x^2(p) \; x(1)x(2) \; x(2)x(3) \, - - \, . \, . \, . \; x(1)x(p)}_{p' \text{ features}}]$

Polynomial Regression

degree m

Univariate (1-dim) case: $f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m = \mathbf{X}\beta$

d=1

where $\mathbf{X} = [1 \;\; X \;\; X^2 \ldots X^m], \beta = [\beta_1 \ldots \beta_m]^T$

$$\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \text{ or } (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y}$$

$$\widehat{f}_n(X) = \mathbf{X}\widehat{\beta}$$

where $\mathbf{A} = \begin{bmatrix} 1 & X_1 & X_1^2 & \ldots & X_1^m \\ \vdots & & & \ddots & \vdots \\ 1 & X_n & X_n^2 & \ldots & X_n^m \end{bmatrix}$

original linear

$A = \begin{bmatrix} 1 & X_1 \\ \vdots & \\ 1 & X_n \end{bmatrix}$

Multivariate (p-dim) case:

$$f(X) = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \cdots + \beta_p X^{(p)}$$
$$+ \sum_{i=1}^{p}\sum_{j=1}^{p} \beta_{ij} X^{(i)} X^{(j)} + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} X^{(i)} X^{(j)} X^{(k)}$$
$$+ \ldots \text{terms up to degree m}$$

Polynomial Regression

blue function

$y_i = f^*(x_i) + \varepsilon_i$

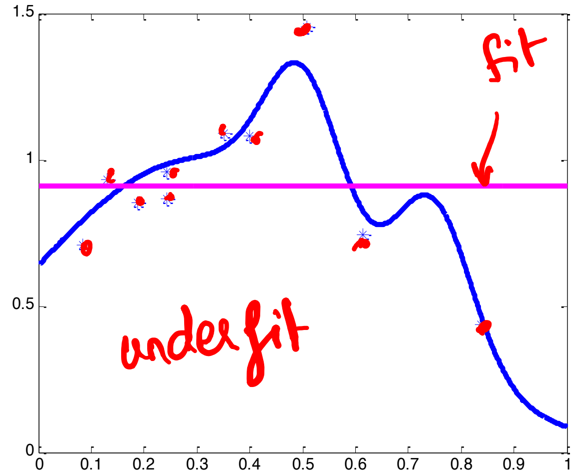
$E[(y - f(x))^2]$

$\frac{1}{n}\sum_{i=1}^{n}$

Polynomial of order k, equivalently of degree up to k-1

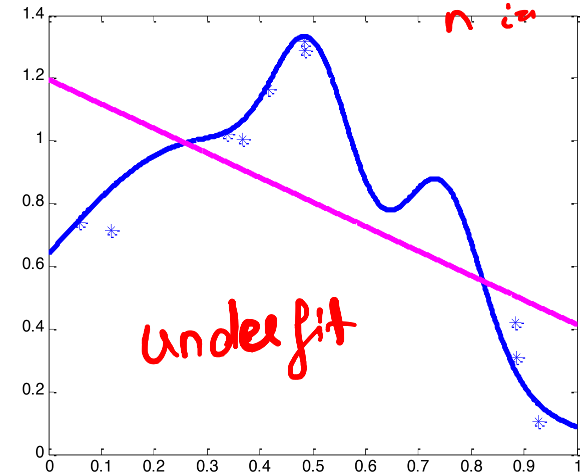
k=1

constant fit



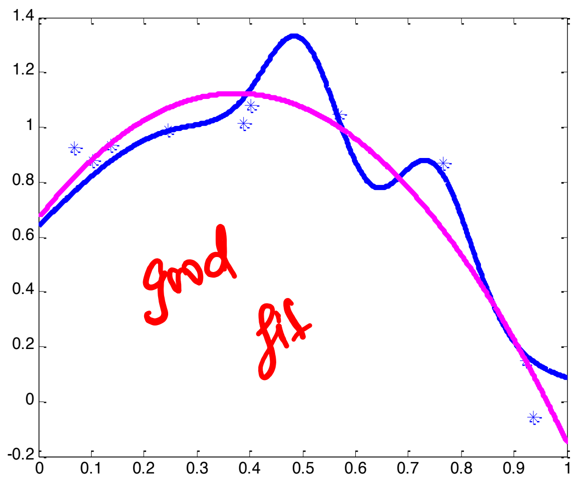
k=2

linear fit



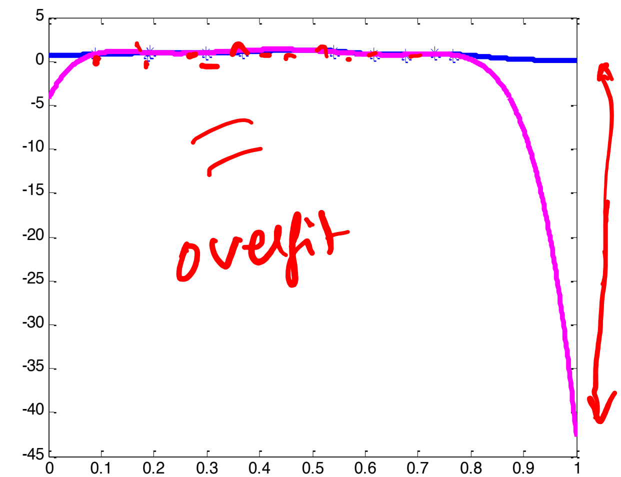
k=3

quadratic fit



k=7

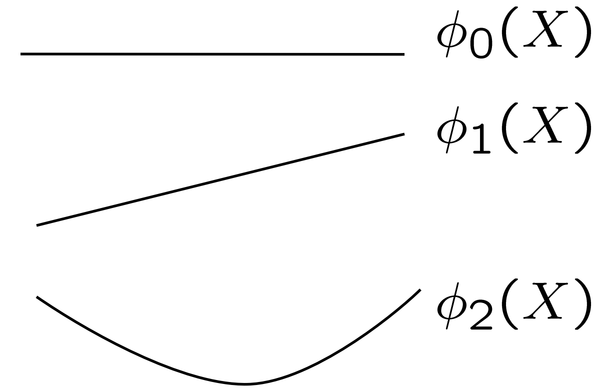
sixth order poly fit



Regression with nonlinear features

$$f(X) = \sum_{j=0}^{m} \beta_j X^j = \sum_{j=0}^{m} \beta_j \underbrace{\phi_j(X)}_{\text{Nonlinear features}}$$

Weight of each feature



In general, use any nonlinear features

e.g. e^X, log X, 1/X, sin(X), …

$$\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y}$$

or

$$(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y}$$

$$\mathbf{A} = \begin{bmatrix} \phi_0(X_1) & \phi_1(X_1) & \dots & \phi_m(X_1) \\ \vdots & & \ddots & \vdots \\ \phi_0(X_n) & \phi_1(X_n) & \dots & \phi_m(X_n) \end{bmatrix}$$

$$\widehat{f}_n(X) = \mathbf{X}\widehat{\beta}$$

$$\mathbf{X} = [\phi_0(X)\ \phi_1(X)\ \dots\ \phi_m(X)]$$