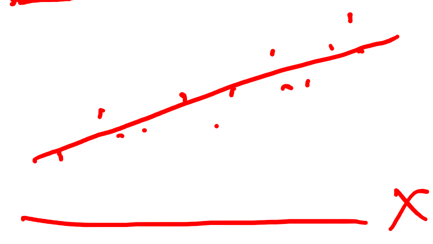
Linear Regression (Matrix-vector form)

$X = [1 \; X_1 \; X_2 \, . \, . \, X_p]$

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

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



$$= \arg\min_{\beta} \frac{1}{n} \underbrace{(\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y})}_{J(\beta)}$$

$$\underset{n\times p}{\mathbf{A}} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}$$

Normal equations: $\underbrace{(\mathbf{A}^T\mathbf{A})}_{p\times p}\widehat{\beta} = \mathbf{A}^T\mathbf{Y}$

If invertible, closed form expression or gradient descent

Linear regression solution satisfies Normal Equations

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

p x p p x1 p x1

$\boxed{A_{n\times p}}$

When is $(\mathbf{A}^T\mathbf{A})$ invertible ?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T\mathbf{A})$?

If $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$, then

$n \times p$ S - r x r

$$\underset{n\times r}{\left[\quad\right]}\ \underbrace{\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_r\end{bmatrix}}_{S}\ \underset{r\times p}{\left[\quad\right]}$$

$U^TU = I = V^TV$

normal equations $(\mathbf{S}\mathbf{V}^\top)\hat{\beta} = (\mathbf{U}^\top\mathbf{Y})$

r x p p x 1 r x 1

$\cancel{V^T}(\cancel{V}SU^TUSV^T)\hat\beta = \cancel{V^T}\cancel{V}SU^TY$

$r \le \min(n,p)$

r equations in p unknowns. Under-determined if r < p, hence no unique solution.

$n < p$ high-dim setting

Regularized Linear Regression

Aarti Singh & Geoff Gordon

Machine Learning 10-701
Feb 22, 2021

Regularized Least Squares

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

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

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

Ridge Regression
(l2 penalty)

$$= \arg\min_{\beta} \quad (\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) \; + \lambda\|\beta\|_2^2$$

$\lambda \geq 0$

→ $\beta^T(A^TA+\lambda I)\beta$ $\quad \beta^T A^T A \beta$ $\quad \beta^T\beta$

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

$(A^TA+\lambda I)v = \lambda' v$

$A^TAv + \lambda v = \lambda' v$

Is $(\mathbf{A}^\top\mathbf{A} + \lambda\mathbf{I})$ invertible ?

Understanding regularized Least Squares

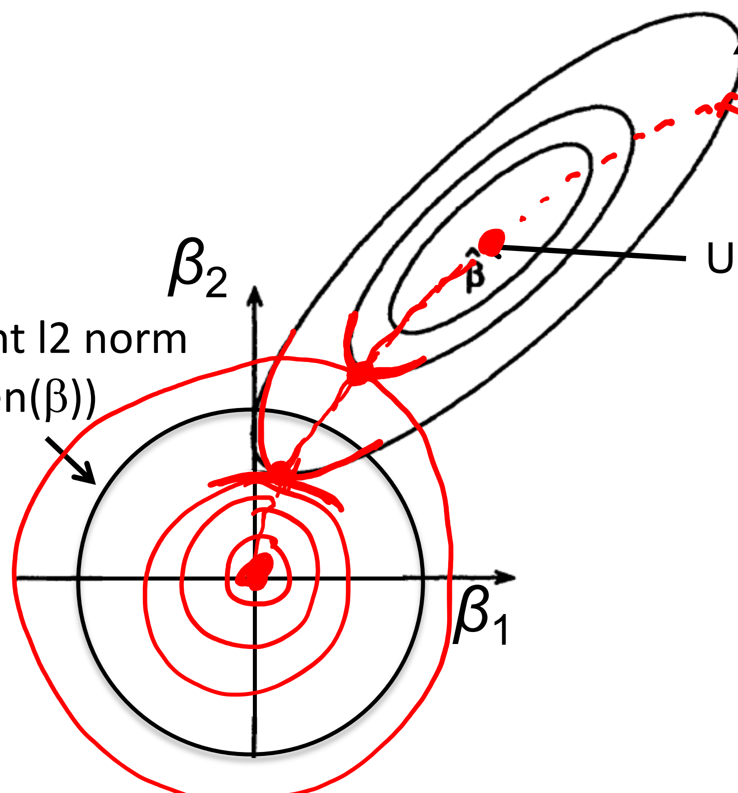
$$\min_\beta (\mathbf{A}\beta - \mathbf{Y})^T(\mathbf{A}\beta - \mathbf{Y}) + \lambda \text{pen}(\beta) = \min_\beta \underbrace{J(\beta) + \boxed{\lambda}\text{pen}(\beta)}_{J^*}$$

$\lambda \geq 0$

Ridge Regression:

$$\text{pen}(\beta) = \|\beta\|_2^2 = \beta_1^2 + \beta_2^2$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$



Regularized Least Squares

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

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

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

$= \sum_i \beta_i^2$

Ridge Regression
(l2 penalty)

$\lambda \geq 0$

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

$= \sum_i |\beta_i|$

Lasso
(l1 penalty)

Many β can be zero – many inputs are irrelevant to prediction in high-dimensional settings (typically intercept term not penalized)

Regularized Least Squares

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

r equations , p unknowns – underdetermined system of linear equations
many feasible solutions

Need to constrain solution further

e.g. bias solution to "small" values of β (small changes in input don't translate to large changes in output)

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

β^2

Ridge Regression
(l2 penalty)

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

$|\beta|$

$\lambda \geq 0$

Lasso
(l1 penalty)

No closed form solution, but can optimize using sub-gradient descent (packages available)

Ridge Regression vs Lasso

$$\min_{\beta} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y}) + \lambda \text{pen}(\beta) = \min_{\beta} J(\beta) + \lambda \text{pen}(\beta)$$

$\|\beta\|_0 = \sum_i 1_{\beta_i \neq 0}$

non-convex

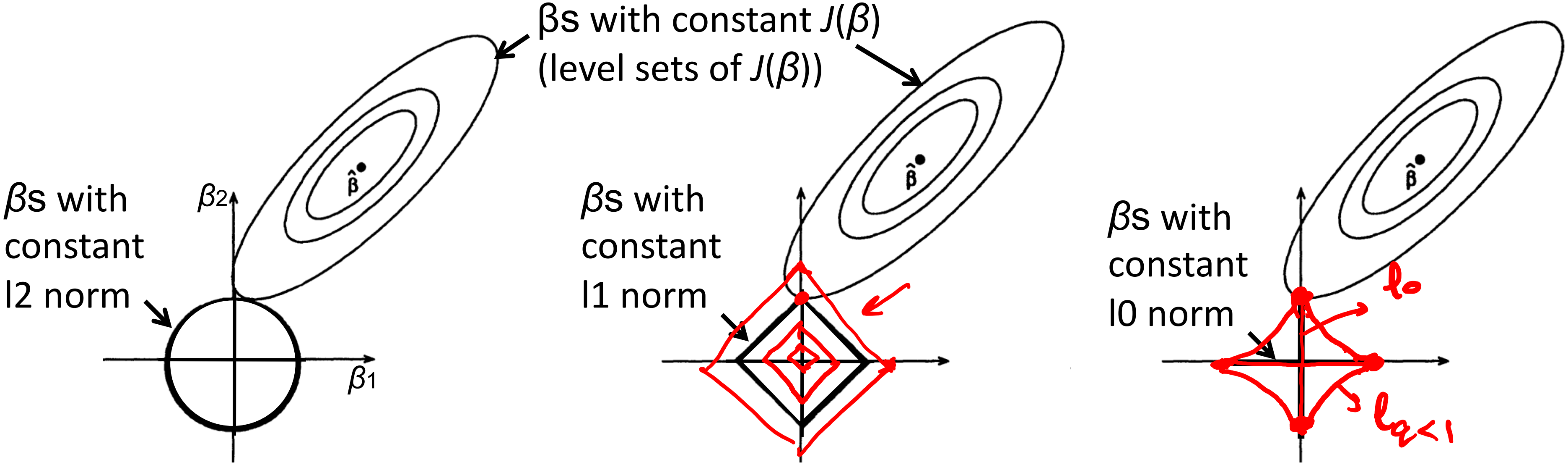
Ridge Regression:

$\text{pen}(\beta) = \|\beta\|_2^2$

Lasso:

$\text{pen}(\beta) = \|\beta\|_1 = \sum_i |\beta_i|$

Ideally l0 penalty, but optimization becomes non-convex



Lasso (l1 penalty) results in sparse solutions – vector with more zero coordinates
Good for high-dimensional problems – don't have to store all coordinates, interpretable solution!

Matlab example

```
clear all
close all

n = 80;     % datapoints
p = 100;   % features
k = 10;     % non-zero features

rng(20);
X = randn(n,p);
weights = zeros(p,1);
weights(1:k) = randn(k,1)+10;
noise = randn(n,1) * 0.5;
Y = X*weights +  noise;

Xtest = randn(n,p);
noise = randn(n,1) * 0.5;
Ytest = Xtest*weights + noise;
```

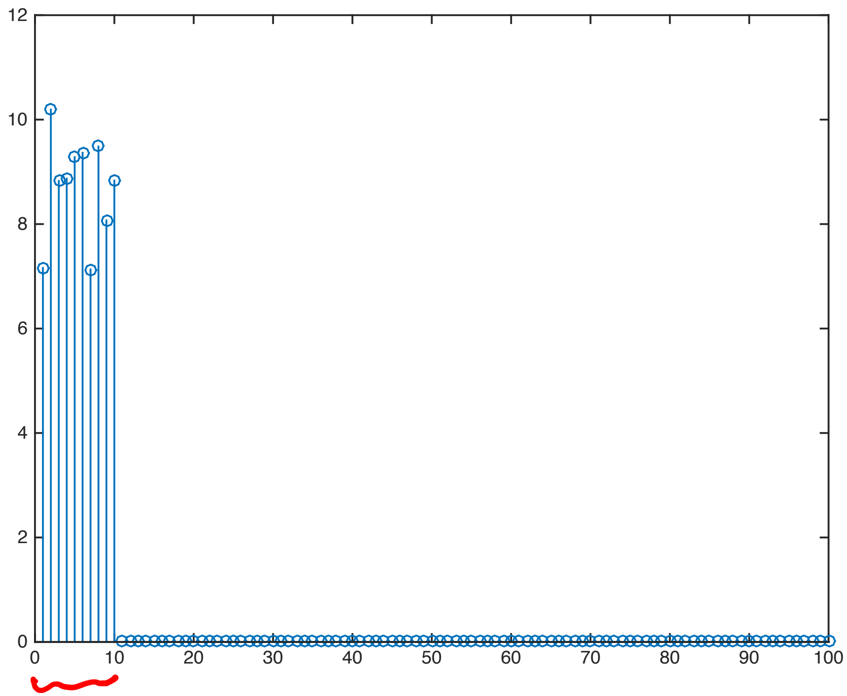
```
lassoWeights = lasso(X,Y,'Lambda',1,
'Alpha', 1.0);
Ylasso = Xtest*lassoWeights;
norm(Ytest-Ylasso)

ridgeWeights = lasso(X,Y,'Lambda',1,
'Alpha', 0.0001);
Yridge = Xtest*ridgeWeights;
norm(Ytest-Yridge)

stem(lassoWeights)
pause
stem(ridgeWeights)
```

Matlab example

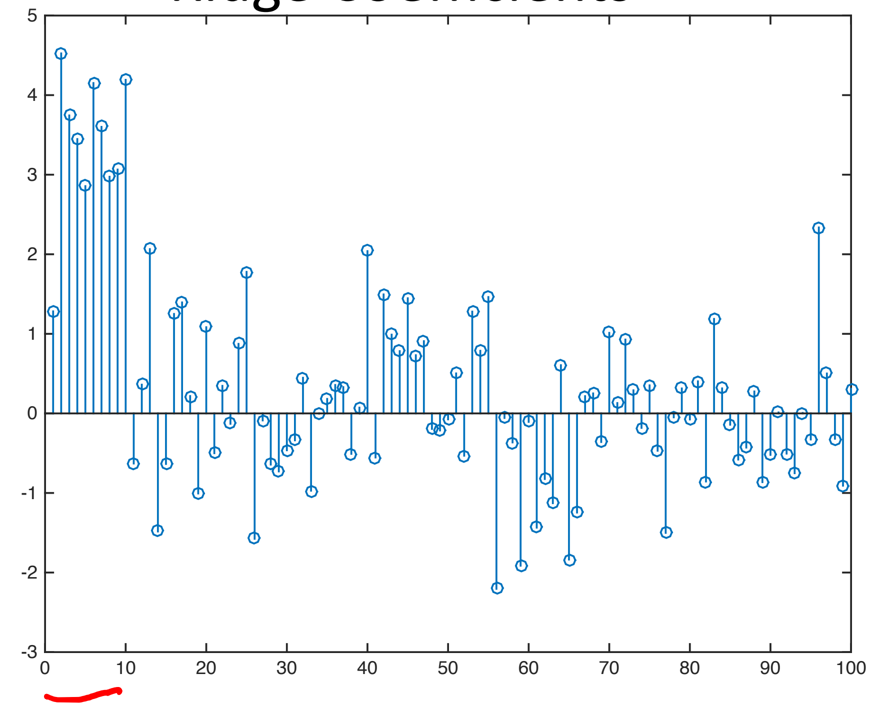
Test MSE = 33.7997

Test MSE = 185.9948

Lasso Coefficients

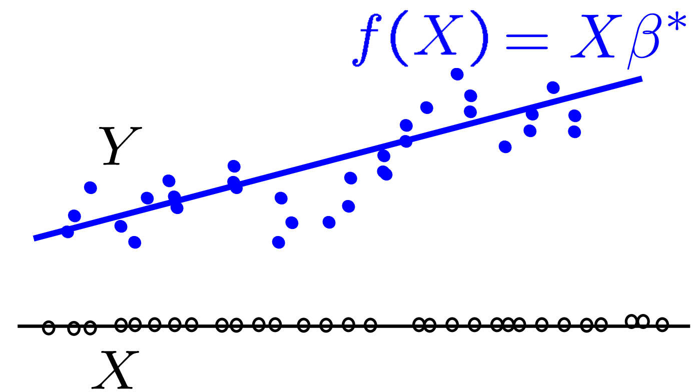


Ridge Coefficients



Least Squares and M(C)LE

Intuition: Signal plus (zero-mean) Noise model



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

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \qquad \underset{1\times}{Y} \sim \mathcal{N}(X\beta^*, \sigma^2)$$

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

$$= \arg\max_{\beta} \log e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}}$$

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

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

Regularized Least Squares and M(C)AP

$P(\theta|D) \propto P(\theta)\,P(D|\theta)$

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

$\log p(D|\theta) \qquad + \log p(\theta)$

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

I) Gaussian Prior

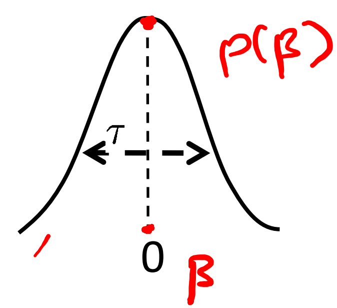
Σ

$$\beta \sim \mathcal{N}(0, \tau^2\mathbf{I})$$

$\beta^T\Sigma^{-1}\beta$

$$p(\beta) \propto e^{-\beta^T\beta/2\tau^2}$$

$\|\beta\|^2$



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

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

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

Ridge Regression

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

Regularized Least Squares and M(C)AP

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

$-\sum_i |\beta_i| = -\|\beta\|_1$

$=$

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

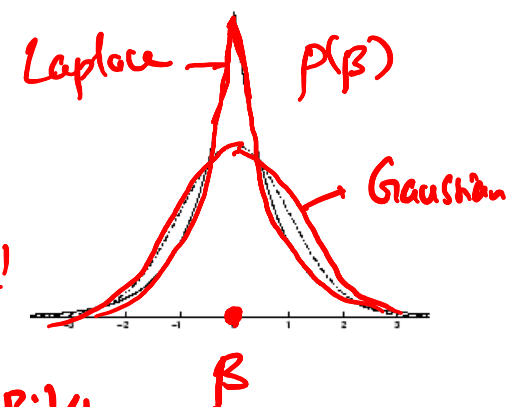
$p(\beta) \propto e^{-|\beta|}$

II) Laplace Prior

$$\beta_i \overset{iid}{\sim} \text{Laplace}(0, t) \qquad p(\beta_i) \propto e^{-|\beta_i|/t}$$

$p(\beta) = \prod_i p(\beta_i) = \prod_i e^{-\frac{|\beta_i|}{t}}$

$= e^{-\sum_i |\beta_i|/t}$



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

Lasso

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

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

Polynomial Regression

$X = [1 \;\; x \;\; x^2 \;\; x^3 \cdots x^m]$

degree m

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

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

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

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

Multivariate (p-dim) case:

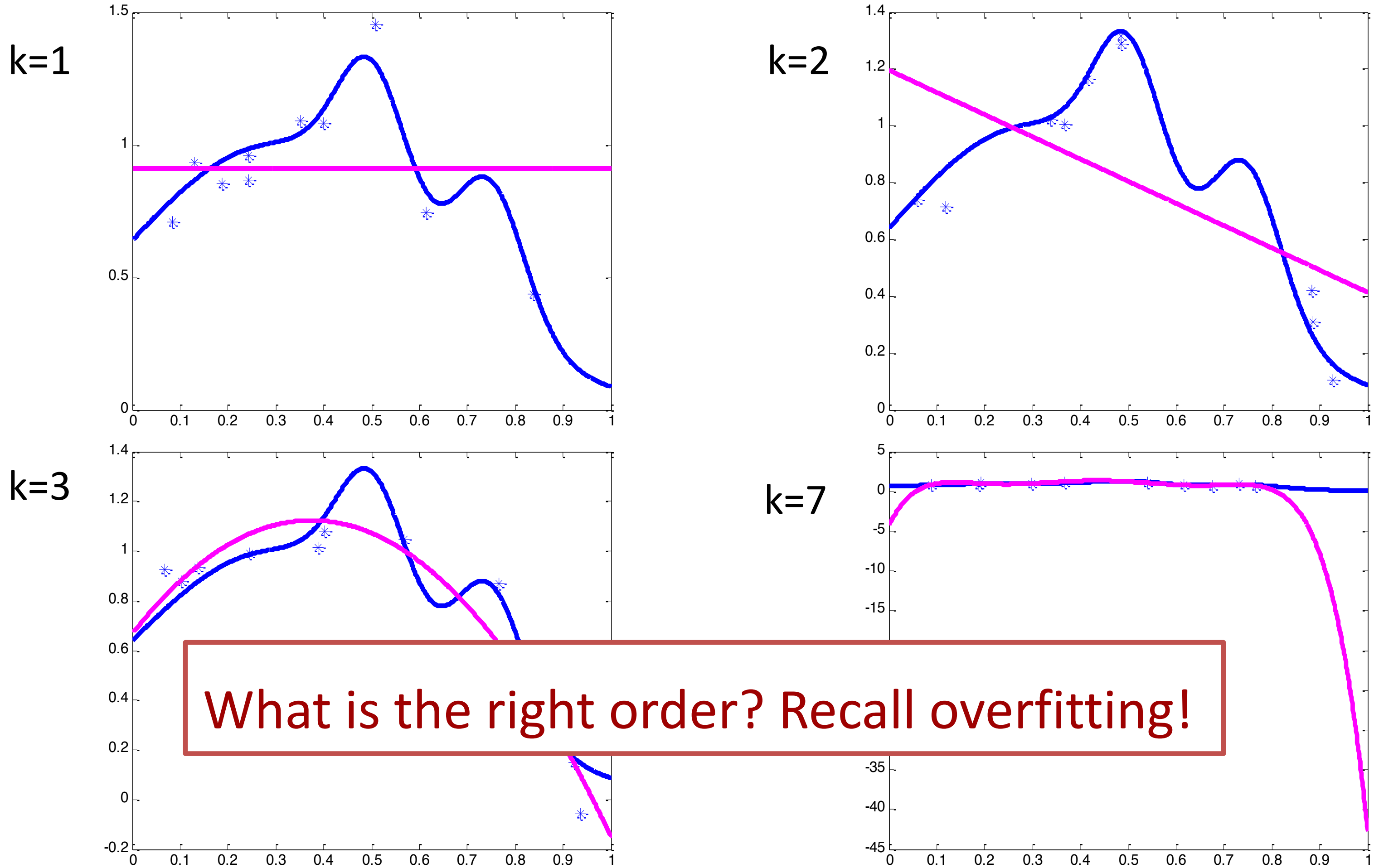
$$f(X) = \beta_0 + \beta_1 X^{(1)} + \beta_2 X^{(2)} + \cdots + \beta_p X^{(p)} + \sum_{i=1}^{p}\sum_{j=1}^{p} \beta_{ij} X^{(i)} X^{(j)} + \sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p} X^{(i)} X^{(j)} X^{(k)} + \ldots \text{terms up to degree m}$$

$p = 2$

$x \to [1 \;\; x_1 \;\; x_2 \;\; x_1^2 \;\; x_2^2 \;\; x_1 x_2]$

Polynomial Regression

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

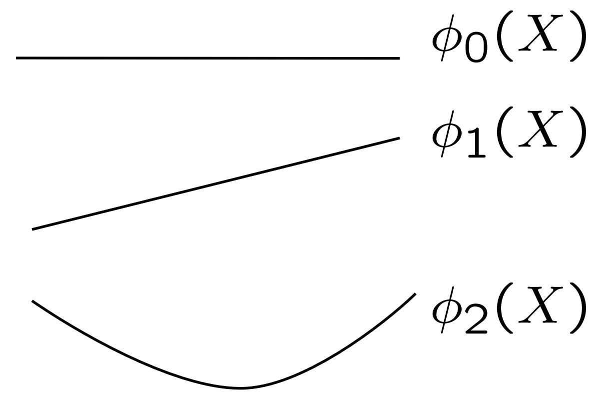


What is the right order? Recall overfitting!

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), …

$$\widehat{\beta} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{Y} \qquad \mathbf{A} = \begin{bmatrix} \phi_0(X_1) & \phi_1(X_1) & \dots & \phi_m(X_1) \\ \vdots & & \ddots & \vdots \\ \phi_0(X_n) & \phi_1(X_n) & \dots & \phi_m(X_n) \end{bmatrix}$$

$$\widehat{f}_n(X) = \mathbf{X}\widehat{\beta} \qquad \mathbf{X} = [\phi_0(X)\ \phi_1(X)\ \dots\ \phi_m(X)]$$