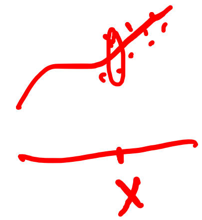
Regularized Linear Regression

Aarti Singh

Machine Learning 10-701

Mar 22, 2023

Mean square error regression



Optimal predictor: $$f^* = \arg\min_{f} \mathbb{E}_{XY}[(f(X) - Y)^2]$$

$= E[Y|X]$

$P(X,Y)$ known

Empirical Minimizer: $$\widehat{f}_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - Y_i)^2$$

$\{X_i, Y_i\}_{i=1}^{n} \overset{iid}{\sim} P(X,Y)$

Class of predictors

$\mathcal{F}$ - Class of Linear functions

- Class of Polynomial functions
- Class of nonlinear functions

$f(X) = X\beta$

$= [X^{(1)} \cdots X^{(p)}] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$

Least Square solution satisfies Normal Equations

$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}$

$A = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & & \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix}_{n \times p}$

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

p x p p x1 p x1

If $(\mathbf{A}^T\mathbf{A})$ is invertible,

1) If dimension p not too large, analytical solution:

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

2) If dimension p is large, computing inverse is expensive O(p^3)
Gradient descent since objective is convex (**A**^{T}**A** $\succeq$ 0)

$$\boxed{\begin{aligned} \beta^{t+1} &= \beta^t - \frac{\alpha}{2}\frac{\partial J(\beta)}{\partial \beta}\Big|_t \\ &= \beta^t - \alpha\ \mathbf{A}^T(\mathbf{A}\beta^t - Y) \end{aligned}}$$

Linear regression solution satisfies Normal Equations

$$A = \begin{bmatrix} x_1^{(1)} & \cdots & x_1^{(p)} \\ \vdots & & \vdots \\ x_n^{(1)} & \cdots & x_n^{(p)} \end{bmatrix}_{n \times p}$$

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

p x p p x1 p x1

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

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

$$A_{n\times p} = \underset{n\times r}{U}\ \underset{r\times r}{S}\ \underset{r\times p}{V^T}$$

rank$(A) = r \le p$ $r \le \min(n,p)$

$$S = \begin{bmatrix} s_1 & & 0 \\ & \ddots & \\ 0 & & s_r \end{bmatrix}$$

$$A^TA = (USV^T)^T USV$$

$$= VS\cancel{U^TU}SV = \underbrace{VS^2V^T}$$

$U^TU = I$

$$\Rightarrow \text{eig}(A^TA) = s_1^2 \ldots s_r^2$$

if $n < p$ (n: no. of datapoints, p: no. of features)
then A^TA is not invertible

Linear regression solution satisfies Normal Equations

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

p x p   p x1   p x1

p equations in p unknowns ($\hat{\beta}$)

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

S - r x r

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

r x p   p x 1   r x 1

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

Regularized Least Squares

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ $A = \begin{bmatrix} X_1^{(1)} & \cdots & X_1^{(p)} \\ X_n^{(1)} & \cdots & X_n^{(p)} \end{bmatrix}$

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)

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

$\lambda \geq 0$

$2A^TA\beta - 2A^TY + 2\lambda\beta = 0$

Ridge Regression

Ridge Regression
(l2 penalty)

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

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

$\lambda \geq 0$

$$= \quad 2A^TA\beta - 2A^TY + \underbrace{2\lambda\beta}_{2\lambda I\beta}$$

$$= \quad 2(A^TA + \lambda I)\beta - 2A^TY \quad = 0$$

$\Sigma = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \ddots \, \lambda_p \end{bmatrix}$

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

$M + \lambda I$

$= V\Sigma V^T + \lambda VV^T$

$V(\Sigma + \lambda I)V^T$

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

always $\lambda > 0$

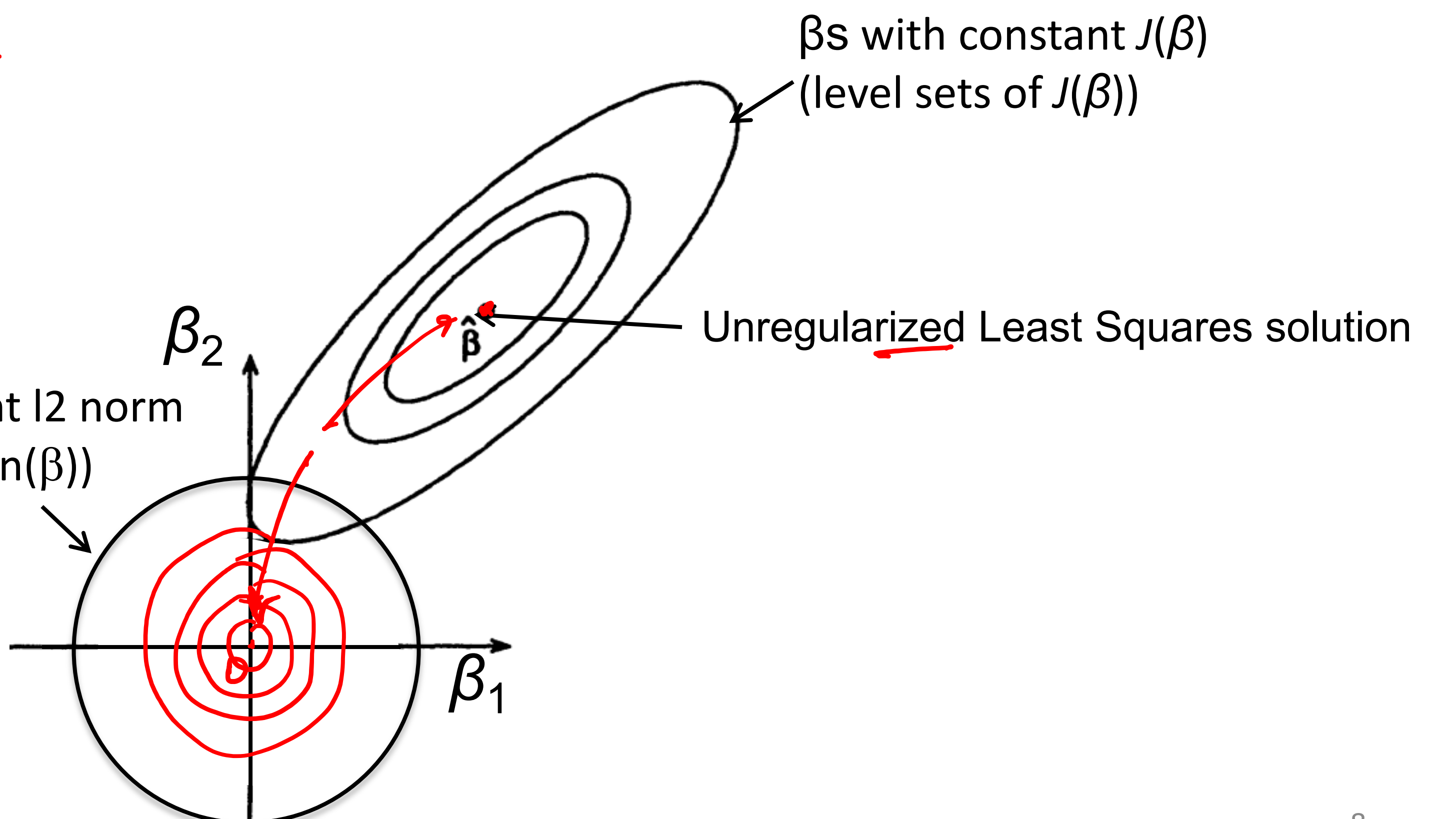
$\text{eig}(M + \lambda I) = \text{eig}(M) + \lambda \geq \lambda > 0$

Understanding regularized Least Squares

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

Ridge Regression:

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



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)

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

Ridge Regression
(l2 penalty)

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

$\lambda \geq 0$

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)

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

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

Ridge Regression
(l2 penalty)

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

$\lambda \geq 0$

Lasso
(l1 penalty)

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

Ridge Regression vs Lasso

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

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

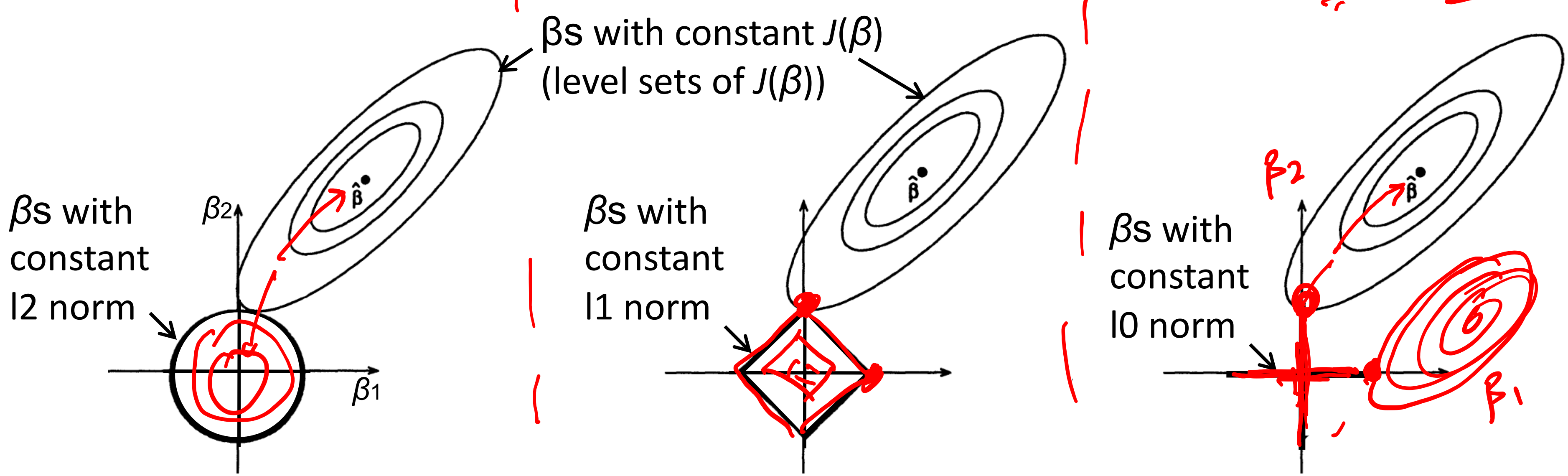
Ridge Regression:

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

Lasso:

$\mathrm{pen}(\beta) = \|\beta\|_1$

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<p

```
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;
```

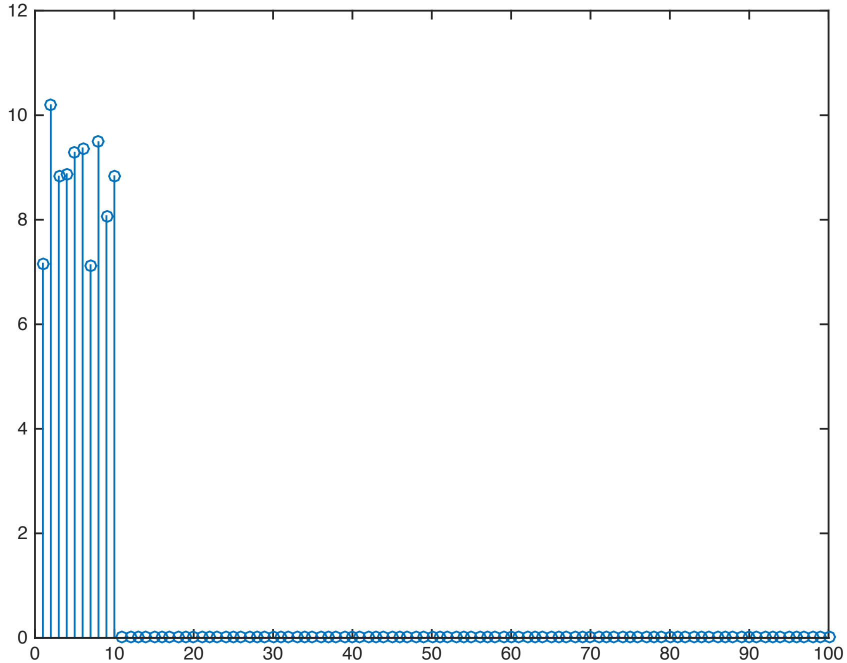
```
lassoWeights = lasso(X,Y,'Lambda',1,
'Alpha', 1.0);
Ylasso = Xtest*lassoWeights;
norm(Ytest-Ylasso)

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

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

Matlab example

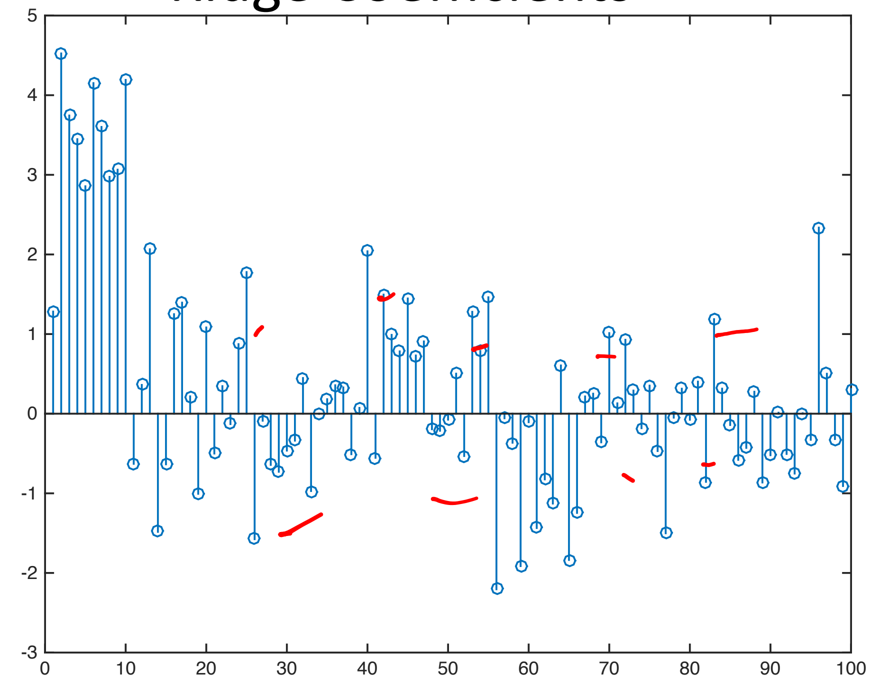
Test MSE = 33.7997

Test MSE = 185.9948

Lasso Coefficients



Ridge Coefficients

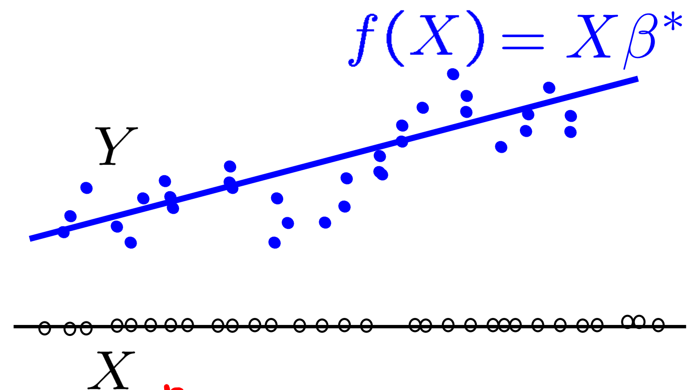


$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2$$

Least Squares and M(C)LE

$E[Y|X]$ ✓

Intuition: Signal plus (zero-mean) Noise model



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

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

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

$$= \arg\max_{\beta} \sum_{i=1}^n \log \underbrace{\frac{1}{(\sqrt{2\pi\sigma^2})^d}}_{a} \underbrace{\exp\left(-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}\right)}_{b}$$

$$\log(ab) = \log a + \log b$$

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

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

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 ?

$p(\beta)$ - prior

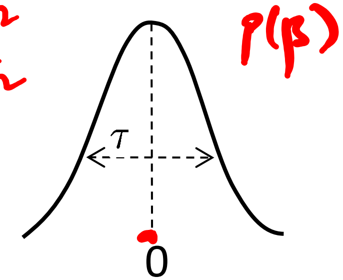
$p(\beta|\text{Data}) \propto p(\text{Data}|\beta) \cdot p(\beta)$

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

I) Gaussian Prior

$\log p(\beta) \propto -\frac{\beta^T\beta}{2\tau^2} = \frac{-\|\beta\|^2}{2\tau^2}$

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



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

Ridge Regression

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

Regularized Least Squares and M(C)AP

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

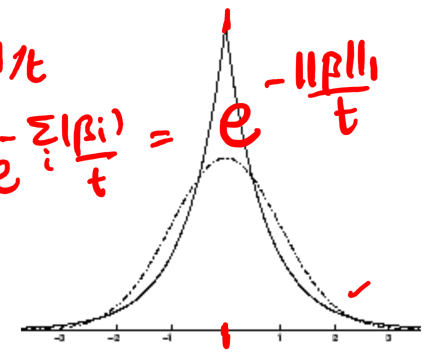
$$\widehat{\beta}_{\text{MAP}} = \arg\max_{\beta} \underbrace{\log p(\{Y_i\}_{i=1}^n|\beta,\sigma^2,\{X_i\}_{i=1}^n)}_{\text{Conditional log likelihood}} + \underbrace{\log p(\beta)}_{\text{log prior}}$$

$= \|\beta\|_1$

II) Laplace Prior

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

$p(\beta) = \prod_i p(\beta_i) \propto \prod_i e^{-|\beta_i|/t} = e^{-\sum_i |\beta_i|/t} = e^{-\|\beta\|_1/t}$



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

Lasso

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

Polynomial Regression

degree m

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

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

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

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

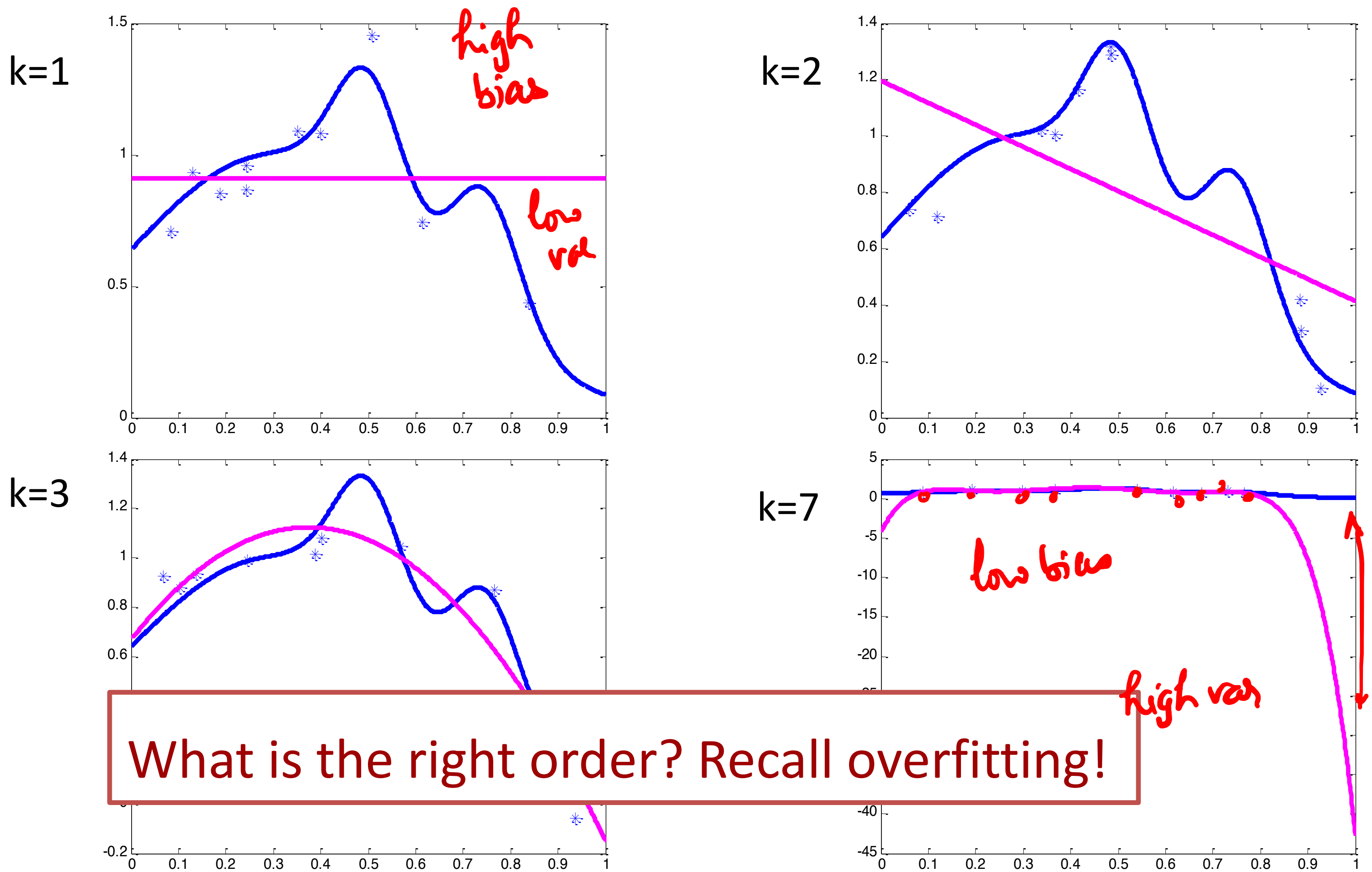
where $\mathbf{A} = \begin{bmatrix} 1 & X_1 & X_1^2 & \ldots & X_1^m \\ \vdots & & & \ddots & \vdots \\ 1 & X_n & X_n^2 & \ldots & X_n^m \end{bmatrix}$

Multivariate (p-dim) case:

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

Polynomial Regression

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

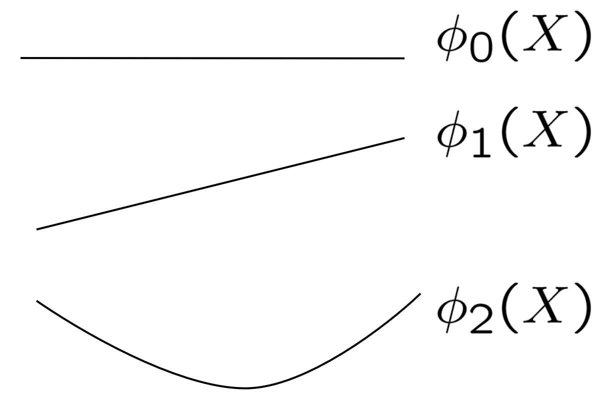


What is the right order? Recall overfitting!

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), …

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

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

Poll

- The maximum likelihood estimate of model parameter α for the random variable $y \sim N(\alpha\, x_1 x_2^3, \sigma^2)$, where x_1 and x_2 are random variables, can be learned using linear regression on n iid samples of (x_1,x_2,y)
  - True
  - False

$y \sim N(\alpha x_1 x_2, \sigma^2)$ ✓

$y \sim N(\alpha x_1 + x_2, \sigma^2)$ ✓

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \rightarrow x_1 x_2^3 = z$

$y \sim N(\alpha z, \sigma^2)$

Can we kernelize linear regression?

Linear (Ridge) regression

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2 \qquad \widehat{\beta} = (\underbrace{\mathbf{A}^T\mathbf{A}}_{}^{p \times p} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y}$$

Recall

$$\mathbf{A} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix}$$

$x_i \cdot x_j = x_i^T x_j$

$\phi(x_i) \cdot \phi(x_j)$

$K(x_i, x_j)$

Hence $\mathbf{A}^T\mathbf{A}$ is a p x p matrix whose entries denote the (sample) correlation between the features

NOT inner products between the data points – the inner product matrix would be $\mathbf{A}\mathbf{A}^T$ which is n x n (also known as Gram matrix)

Using dual formulation, we can write the solution in terms of $\mathbf{A}\mathbf{A}^T$

Ridge regression

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2 \qquad \widehat{\beta} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y}$$

<u>Similarity with SVMs</u>

Primal problem:

$$\min_{\beta, z_i} \sum_{i=1}^{n} z_i^2 + \lambda\|\beta\|_2^2$$
$$\text{s.t. } z_i = Y_i - X_i\beta \qquad \alpha_i$$

SVM Primal problem:

$$\min_{w, \xi_i} C\sum_{i=1}^{n} \xi_i + \frac{1}{2}\|w\|_2^2$$
$$\text{s.t. } \xi_i = \max(1 - Y_i X_i w, 0)$$

Lagrangian:

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

α_i – Lagrange parameter, one per training point

Ridge regression (dual)

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - X_i\beta)^2 + \lambda\|\beta\|_2^2 \qquad \widehat{\beta} = (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y}$$

Dual problem:

$$\max_{\alpha} \min_{\beta, z_i} \sum_{i=1}^{n} z_i^2 + \lambda\|\beta\|^2 + \sum_{i=1}^{n} \alpha_i(z_i - Y_i + X_i\beta)$$

$\alpha = \{\alpha_i\}$ for i = 1,…, n

Taking derivatives of Lagrangian wrt β and z_i we get:

$$\beta = -\frac{1}{2\lambda}\mathbf{A}^\top\alpha \qquad z_i = -\frac{\alpha_i}{2}$$

Dual problem: $$\max_{\alpha} \; -\frac{\alpha^\top\alpha}{4} - \frac{1}{4\lambda}\alpha^\top\mathbf{A}\mathbf{A}^\top\alpha - \alpha^\top\mathbf{Y}$$

n-dimensional optimization problem

Ridge regression (dual)

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

$$\begin{aligned}\widehat{\beta} &= (\mathbf{A}^T\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{Y} \\ &= \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \lambda\mathbf{I})^{-1}\mathbf{Y}\end{aligned}$$

Dual problem:

$$\max_\alpha \; -\frac{\alpha^\top\alpha}{4} - \frac{1}{4\lambda}\alpha^\top\mathbf{A}\mathbf{A}^\top\alpha - \alpha^\top\mathbf{Y} \qquad \Rightarrow \widehat{\alpha} = -\left(\frac{\mathbf{A}\mathbf{A}^\top}{\lambda} + \mathbf{I}\right)^{-1} 2\,\mathbf{Y}$$

can get back $\hat{\beta} = -\frac{1}{2\lambda}\mathbf{A}^\top\hat{\alpha} \;\; = \mathbf{A}^\top(\mathbf{A}\mathbf{A}^\top + \lambda\mathbf{I})^{-1}\mathbf{Y}$

Weighted average of training points

Weight of each training point (but typically not sparse)

Kernelized ridge regression

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

Using dual, can re-write solution as:

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

How does this help?

- Only need to invert n x n matrix (instead of p x p or m x m)
- More importantly, kernel trick!

AA^T involves only inner products between the training points
BUT still have an extra **A**^{T}

Recall the predicted label is $\widehat{f}_n(X) = \mathbf{X}\widehat{\beta}$

$$= \underbrace{\mathbf{X}\mathbf{A}^T}_{K_{x,x_i}}(\underbrace{\mathbf{A}\mathbf{A}^T}_{K_{x_i,x_i}} + \lambda\mathbf{I})^{-1}\mathbf{Y}$$

XA^T contains inner products between test point **X** and training points!

Kernelized ridge regression

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

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

Using dual, can re-write solution as:

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

How does this help?

- Only need to invert n x n matrix (instead of p x p or m x m)
- More importantly, kernel trick!

$$\widehat{f}_n(X) = \mathbf{K}_X(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{Y}$$ where $$\mathbf{K}_X(i) = \boldsymbol{\phi}(X)\cdot\boldsymbol{\phi}(X_i)$$ $$\mathbf{K}(i,j) = \boldsymbol{\phi}(X_i)\cdot\boldsymbol{\phi}(X_j)$$

Work with kernels, never need to write out the high-dim vectors

Ridge Regression with (implicit) nonlinear features $\boldsymbol{\phi}(X)$! $f(X) = \phi(X)\beta$