

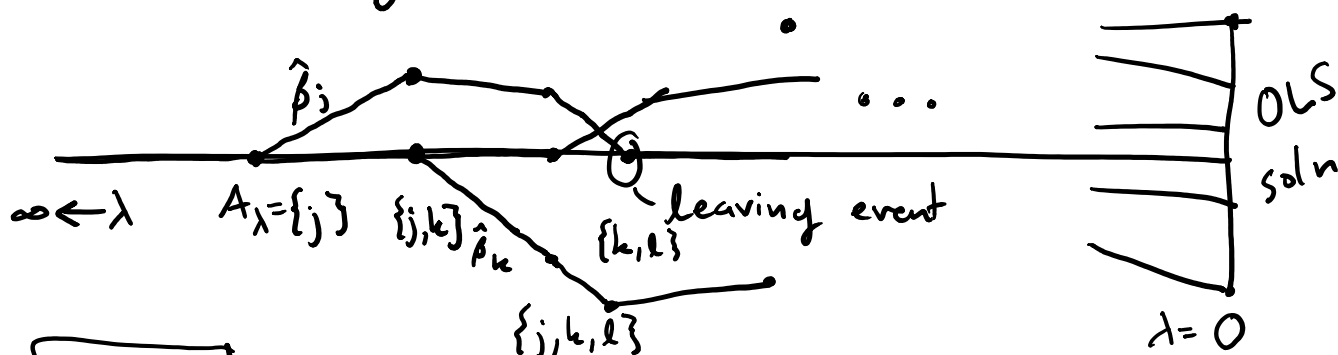
# More about the Lasso

Wednesday, April 19, 2017 5:47 PM

Recall the lasso:  $\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$ ,

Define  $A_\lambda = \text{supp}(\hat{\beta}_\lambda) = \{j: \hat{\beta}_{\lambda j} \neq 0\}$ .

As  $\lambda \rightarrow \infty$ ,  $\beta \rightarrow 0$  so start at  $\lambda = \infty$  and eventually as  $\lambda$  decreases  $A_\lambda \neq \{\}$ .



**fact 1**  $\hat{\beta}_j$  is continuous, piecewise linear in  $\lambda$

**fact 2** Slope of  $\hat{\beta}_j$  is regression coef on residual for OLS on  $A_\lambda$ .

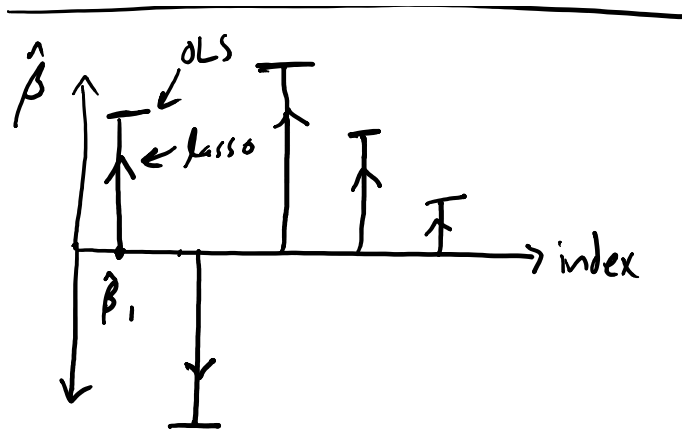
**fact 3** LAR with lasso modification solves the lasso.

We can think of the lasso path as a sequence of models  $A_{\lambda_1}, A_{\lambda_2}, \dots$  for  $\lambda_1 > \lambda_2 \dots$  knots.

Lasso introduces a bias

$$\hat{\beta} \uparrow \downarrow^{OLS} T$$

So, for selected  
model 1 1 1 .



model A, solve  
restricted OLS,  

$$\tilde{\beta} = (X_A^T X_A)^{-1} X_A^T y$$

Clarification:  $\beta_j = \beta_{j+} - \beta_{j-}$  for  $\beta_{j+}, \beta_{j-} \geq 0$

$$\text{QP: } \min_{\beta_+, \beta_-} \|y - X(\beta_+ - \beta_-)\|_2^2 \text{ s.t. } \sum_j \beta_{j+} + \beta_{j-} \leq C$$

$$\beta_{j+}, \beta_{j-} \geq 0$$

if  $\beta_+, \beta_-$  feasible then  $\sum_j |\beta_j| \leq \sum_j \beta_{j+} + \beta_{j-} \leq C$   
 for  $\beta := \beta_+ - \beta_- \Rightarrow \text{QP} \geq \text{Lasso}$

if  $\beta$  is Lasso sol<sup>n</sup>  $\beta_{j+} = (\beta_j)_+, \beta_{j-} = (-\beta_j)_+$   
 $\sum_j \beta_{j+} + \beta_{j-} = \sum_j |\beta_j| \leq C \Rightarrow \text{QP} \leq \text{Lasso}$

# Logistic Regression

Wednesday, April 19, 2017 5:47 PM

Recall Empirical Risk Minimization :

$$\min_{\beta \in \mathcal{D}} \frac{1}{n} \sum_{i=1}^n l(y_i, x_i; \beta) \quad \text{for some loss, } l.$$

Regression:  $l(y_i, x_i; \beta) = (y_i - x_i^T \beta)^2$ .

Gaussian error model :  $Y = X^T \beta + \varepsilon$ ;

for  $\varepsilon_i \sim N(0, \sigma^2)$  then density is

$$f_{Y|X}(y_i | \beta, x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - x_i^T \beta)^2}$$

$$-2\sigma^2 \log f_{Y|X}(y_i | \beta, x_i) = (y_i - x_i^T \beta)^2 + C$$

Maximum likelihood is empirical risk minimization when the loss is the negative log-likelihood (under ind.).

Logistic model Classification ( $Y \in \{0, 1\}$ )

$Y|X$  is Bernoulli (not much choice there)

but how does  $\mathbb{P}\{Y=1 | X=x\}$  depend on  $x$ ?

The logistic model assumes that

$$\log \frac{\mathbb{P}\{Y=1|X=x\}}{\mathbb{P}\{Y=0|X=x\}} = x^T \beta.$$

Logit function is  $\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$ , so,

$$\text{logit}(\mathbb{P}\{Y=1|X=x\}) = x^T \beta.$$

$$\text{Claim: } \text{logit}^{-1}(a) = \frac{e^a}{1+e^a}$$

$$e^{\text{logit}(p)} = \frac{p}{1-p} \quad \text{so}$$

$$\text{logit}^{-1}(\text{logit}(p)) = \frac{p}{1+\frac{p}{1-p}} = p \quad \checkmark$$

$$\begin{aligned} \mathbb{P}\{Y=1|X=x\} \\ = \frac{e^{x^T \beta}}{1+e^{x^T \beta}} \end{aligned}$$

$$\text{Also, } \mathbb{P}\{Y=0|X=x\} = \frac{1}{1+e^{x^T \beta}} \quad \text{so}$$

$$\log \mathbb{P}\{Y=y|X=x\} = y x^T \beta - \log(1+e^{x^T \beta})$$

and the loss (neg. log-likelihood) is

$$l(y, x; \beta) = -y x^T \beta + \log(1+e^{x^T \beta})$$

note: in ESL they maximize log-likelihood.

Logistic Regression (should be called  
Logistic Classification)

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n -y_i x_i^T \beta + \log(1 + e^{x_i^T \beta})$$

# Fitting Logistic Regression

Wednesday, April 19, 2017 8:47 PM

$$\ell(y, x; \beta) = -y x^T \beta + \log(1 + e^{x^T \beta})$$

$$\frac{\partial}{\partial \beta} \ell(y, x; \beta) = -y x + \frac{e^{x^T \beta} \cdot x}{1 + e^{x^T \beta}} = -y x + \text{logit}^{-1}(x^T \beta) \cdot x$$

$$= (p - y) x \quad \text{if } p = \mathbb{P}\{Y=1 | X=x, \beta\}$$

$$\frac{\partial^2}{\partial \beta \partial \beta^T} \ell(y, x; \beta) = \frac{e^{x^T \beta} x x^T}{1 + e^{x^T \beta}} - \frac{e^{2x^T \beta} x x^T}{(1 + e^{x^T \beta})^2} = \frac{e^{x^T \beta}}{(1 + e^{x^T \beta})^2} x x^T$$

$$= p(1-p) x x^T \geq 0 \quad \text{so } \ell \text{ is convex!}$$

$$\text{Empirical risk: } R_n(\beta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, x_i; \beta)$$

$$\frac{\partial}{\partial \beta} R_n(\beta) = \frac{1}{n} \sum_{i=1}^n (p_i - y_i) x_i = \frac{1}{n} X^T r \quad \text{where}$$

$$p_i = \text{logit}^{-1}(x_i^T \beta), \quad r_i = p_i - y_i$$

$$\frac{\partial^2}{\partial \beta \partial \beta^T} R_n(\beta) = \frac{1}{n} \sum_{i=1}^n p_i(1-p_i) x_i x_i^T = \frac{1}{n} X^T W X$$

$$W_{ii} = p_i(1-p_i)$$



Newton-Raphson

until convergence criteria

$$\beta_{t+1} \leftarrow \beta_t - \underset{\uparrow}{H}^{-1} \underset{\wedge}{g}$$

Hessian  $\hookrightarrow$  gradient at  $\beta_+$

Idea: Approximate  $R_n(\beta)$  by local quadratic

$$R_n(\beta) \approx R_n(\beta_+) + g^T(\beta - \beta_+) + \frac{1}{2}(\beta - \beta_+)^T H(\beta - \beta_+)$$

$$\hookrightarrow \text{argmin} = \beta_+ - H^{-1}g.$$

$$\text{Logistic: } H^{-1}g = \underbrace{(X^T W X)^{-1} X^T r}_{\text{weighted least squares}}$$

Newton Raphson  $\rightarrow$  iteratively re-weighted least squares